Free-Hand Gestures for Music Playback: Deriving Gestures with a User-Centred Process

Niels Henze
Andreas Löcken
Susanne Boll
University of Oldenburg
Oldenburg, Germany
firstname.lastname@uni-oldenburg.de

Tobias Hesselmann
Martin Pielot
OFFIS - Institute for Information Technology
Oldenburg, Germany
firstname.lastname@offis.de

ABSTRACT

Music is a fundamental part of most cultures. Controlling music playback has commonly been used to demonstrate new interaction techniques and algorithm. In particular, controlling music playback has been used to demonstrate and evaluate gesture recognition algorithms. Previous work, however, used gestures that have been defined based on intuition, the developers' preferences, and the respective algorithm's capabilities. In this paper we propose a refined process for deriving gestures from constant user feedback. Along this process a set of free-hand gestures for controlling music playback is developed. The situational context is analyzed to shape the usage scenario and derive an initial set of necessary functions. In a successive user study the set of functions is validated. Furthermore, proposals for gestures are collected from the participants for each function. Two gesture sets containing static and dynamic gestures are derived and analyzed in a comparative evaluation. The evaluation shows that we developed an appropriate set of free-hand gestures for music playback. Our results indicate that the proposed process, that includes validation of each design decision, improves the final results.

Categories and Subject Descriptors

H.5.2 [**Interfaces and Presentation**]: User Interfaces - Interaction styles

General Terms

Design, Human Factors, Experimentation

Keywords

music, gestures, camera, gesture recognition, CD, process

1. INTRODUCTION

Gestural interfaces have been actively explored since the work of Bolt [3]. Those interfaces are often used in movies such Minority Report and show up routinely in popular TV series like CSI: Miami to demonstrate a "futuristic" interaction between users and computers. With the introduction of the Wiimote [20] gestural interfaces became available to a large audience. Computer science literature often motivates gesture based interaction with sentences such as "Gestures are a natural form of communication and are easy to learn" [2]. However, even using gestures to communicate with other humans is a learned communication technique. Furthermore, there is nothing natural about using gestures to control a computer per se.

Controlling music playback (e.g. play, stop, pause, and next) is often used to demonstrate new interfaces and interaction techniques (e.g. [14, 9]). Using a set of function to control music playback has also been used to demonstrate and evaluate gesture recognition algorithms (e.g. [11, 7]). In order to derive meaningful conclusions from an evaluation of a gesture recognition algorithm it is, however, helpful to use a gesture set which is not purely based on the designer's intuition, the algorithms capabilities, or chance. Most work in the area of gestural interaction focused on algorithms and robust recognition of gestures (e.g. [5, 6, 13, 15, 20]). However, gestural interfaces must fulfil the same requirements as any other interaction technique. In particular, it is important to define usable gestures for the functionalities that the particular application offers. In order to deduce usable gestures a process that ensures valid results must be employed.

After discussing the related work in Section 2 we propose a refined process for deriving free-hand gestures from constant user feedback in Section 3. Along this process a set of free-hand gestures for controlling music playback is developed. In Section 4 the situational context is analyzed to shape the usage scenario and derive an initial set of necessary functions. In the successive user study described in Section 5 the set of functions is validated. Furthermore, proposals for gestures are collected from the participants for each function. Two gesture sets containing static and dynamic gestures are derived in Section 6. In Section 7 the gesture sets are and analyzed in a comparative evaluation. Based on the results the gestures are refined to form a consistent set of free-hand gestures for music playback. We close the paper with a conclusion and outlook to future work in Section 8.

2. RELATED WORK

Technically, several approaches can be used for detecting and identifying hand gestures. In [10], we use an infrared camera to detect finger touches and hand gestures on the

MUM '10, December 1-3, 2010, Limassol, Cyprus
Copyright © 2010 ACM 978-1-4503-0424-5/10/12 $10.00.

surface of an interactive tabletop. Schlömer et al. use the accelerometers of a Nintendo Wii controller to detect three-dimensional gestures in open space [20]. Other approaches, including the one presented in this paper, are using ordinary digital cameras to track and identify hands in mid-air. Computer vision algorithms are applied on the sampled images of the camera to recognize hand postures in the picture. Some of the more recent algorithmic approaches include the GUIC system, which combines particle filter algorithms and elastic graph matching [22] and [19] who use a two-layered Bayesian network for hand gesture recognition.

On the application side, several works can be identified which are controlled using mid-air gestures tracked by visual sensors. A product from the industry with similarities to the presented system is Symbian Moove, a gesture-based music player for Symbian S60 based mobile phones [4]. By performing hand gestures in mid-air, users are able to control a music player application that comes with their smartphone. It works by sampling and interpreting pictures from the built-in camera of the mobile device. In contrast to the approach presented in this paper, the system uses a set of four predefined hand gestures: left-to-right and right-to-left motions, covering the camera and tapping the camera. In [21] Stenger et al. present a remote control based on a single camera mounted on a public display, which is controlled by hand gestures. The system is aimed for public settings in particular and can be used to control several applications, including a system to browse video collections as well as viewing a gallery of 3D objects.

Much less research has yet been done in the area of identifying *appropriate* gestures for common and special tasks, and into the investigation of the design space for such gestures. To date, most gesture sets have been defined by system designers (e.g. [5, 6, 13, 15, 20]). Members of those gesture sets are often chosen out of concern for reliable recognition using a technology-based approach. Nielsen et al. claim that, despite skilful design, technology-based approaches lead to an awkward gesture vocabulary without intuitive mapping towards functionality, and a system which works under strictly pre-defined conditions [17]. Kray et al. asked participants to spontaneously produce gestures with their phone to trigger a set of different activities. Their results suggest that phone gestures have the potential to be easily understood by end users and that certain device configurations and activities may be well suited for gesture control [12]. The findings of Morris et al. indicate that users prefer gestures authored by larger groups of people, such as those created by end-user elicitation methodologies or those proposed by more than one researcher [16].

Nielsen et al. proposed a procedure, consisting of three user studies, to derive a usable set of gestures for a given task [17]. They demonstrated the procedure by deriving a gesture set for a simple architectural design application. Likewise Akers employed a similar process to derive a set of gestures for 3D selection of neural pathways [1] and Wobbrock et al. derived basic gestures for surface computing using a participatory approach [23].

3. PROCEDURE

We assume that the procedure defined by Nielsen et al. [17] can be refined by collecting more information from each of the conducted user studies. In particular, we propose to validate the outcome of each user study in the subsequent study. The procedure employed in this paper consists of four steps:

- Usage context and functions
- Participatory design
- Definition of gesture sets
- Evaluation and improvement

In the following we describe and justify each of the four steps.

3.1 Usage context and functions

In first step the usage context of the intended interface is analyzed and an initial set of functions are found. Designers and developers start with a vague idea that a particular application or user interface can be improved by a gestural interface. Thus, understanding the usage context is important not only to derive adequate functionalities. It can also be used to validate the developers' initial idea to design a gestural interface for a particular use case. The designers must concretise the usage scenario in order to design the interface for situations where it makes sense to use gestural input. In particular, location and audience had a significant impact on a user's willingness to perform gestures [18]. Furthermore, a set of functions that is necessary for the application should be collected. In order to support non tech-savvy participants demonstrating a simple prototype can help them to imagine the final system. As no concrete usage scenario is defined beforehand the initial function set is not necessarily conclusive.

3.2 Participatory design

In the second step the initial set of functionalities is validated and multiple gesture sets are derived from participatory design techniques. As the lack of important functionalities can render an interface useless the initial set of functionalities must be validated for the concretized usage scenario. Furthermore, potential gestures for each of the functionalities are collected by conducting a user study and ask the participants to perform gestures for the respective functionalities. These gestures should be recorded on video to analyze the results afterwards. Nielsen et al. highlight that "scenarios take the testees away from technical thinking, especially when conducting the tests on technically minded people" [17]. Likewise we would like to highlight that non tech-savvy participants need a concrete idea about the behaviour of the system. They must be enabled to imagine the situation in which they would use the application in order to avoid socially unacceptable gestures [18].

3.3 Definition of gesture sets

The proposals for gestures must be formalized to define a consistent set of gestures. In contrast to Nielsen et al. [17] we propose to derive multiple gesture sets. By not limiting the outcome to a single set of gestures the risk to reject promising candidates is reduced. Nonetheless, every gesture must be part of a consistent gesture set to ensure that a gesture can be combined with other gestures in a reasonable way. I.e. it must be avoided to define the same gesture or very similar gestures for different functions.

Figure 1: Interviewer and participant during the interview in the participant's apartment.

3.4 Evaluation and improvement

Finally the defined gesture sets must be evaluated and eventually refined. Nielsen et al. conduct a final user study to benchmark the derived gesture set. However, conducting a comparative evaluation of multiple gesture sets enables to compare the respective sets (similar to [18]) as well as the individual gestures. Thereby the designers are not only able to select the gesture set with the higher performance but also to mix gestures from multiple sets if necessary. Furthermore, the evaluation can be used to refine the gestures by collecting qualitative advices from participants.

4. USAGE CONTEXT AND FUNCTIONS

Following the process described above the first step is to concretize the usage context of the intended gesture-based music player and define the initial set of functions. Thus, the goal of this study was understand the situational context in which users listen to music. We intended to concretize in which situations a gestural interface to control music playback should be applied. In addition, we tried to identify the required functionalities to control the music.

4.1 Methodology and Participants

The study was split into two halves. In the first half, we assessed people's music listening goals and needs in different situations through semi-structured interviews. In the second part, we presented them a simple prototype that recognizes CDs using a webcam and plays the according music in order to get initial feedback about desired functionality and behaviour of the system. To not limiting the participants' creativity we did not reveal our intention to design a gestural interface to the participants.

First we asked about situations in which listening to music plays an important role for the interviewed person. For the rest of the interview, one of these situations was picked by the interviewer. Our aim here was to cover a wide range of different situations. In the following part of the interview, more in-depth questions about participant's goals during the respective situations and the role of the music were asked. Then, we investigated which steps the users usually perform while listening to music, what types of music players were used, and how satisfied the participants were with these solutions. Here we aimed at understanding, which functionality of music players would be most important in this situation.

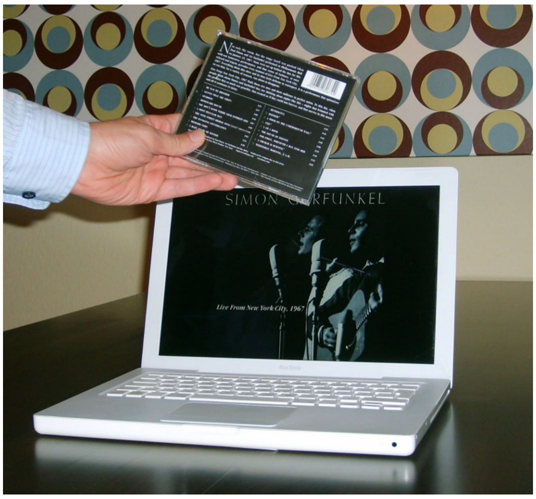

Figure 2: A CD in front of the computer's webcam is recognized and the associated music is started.

After the first part of the interview we presented the prototype described in Section 4.2 to the participants. They could try to select music with the prototype. We then asked them which functionality the prototype should provide in order to be useful to them. We also asked the participants to name good and bad aspects of the intended system.

The interviews took about 20 minutes and were conducted at private places, such as the interview partner's homes as shown in Figure 1. Nine persons from different educational and social backgrounds were interviewed. Their age was between 23 and 32 with a mean of 27.5 (SD: 3.4). Seven of them were male, two female.

4.2 Apparatus

We used a simple prototype to provide a hands-on experience to the participants in the second part of the study. The hardware setup of the prototype consists of a notebook with an integrated webcam and a collection of CDs. On the notebook runs a application that constantly analyses the image stream recorded by a webcam. Using the algorithm described in [8] the prototype recognizes selected album covers. As shown in Figure 2 users can hold an album cover, e.g. a CD's jewel case, in front of the webcam. Using image analysis the corresponding digital music is retrieved and played.

4.3 Results

The interviews were conducted by different interviewers and one participant at a time. During the interview the interviewer took notes using pen and paper and wrote a report about the interview afterwards. Based on these reports the interviewers assembled the following results.

4.3.1 Music listening habits

Situations. We identified three kinds of situations were music was listened to that differ in the meaning of the music. We distinguished between situations were music is the key aspect, music play a major role, and music is secondary.

However, the borders between the three classes are blurred and the meaning of the music in each situation can vary constantly, depending on the current context.

Music was considered the key aspect when the participants reported to listen consciously to it while doing nothing else as a primary task. These situations typically occurred in the home or in the car when driving alone and usually about once a week and if the situation occurs it does not last very long. The second class of situations were those, were music was an integral part of the situation, such as at parties, at work, or in some cases while doing sports. These situations typically occur not daily but last for a longer time then the previous class of situations. The third class of situations were those, were music was listened by the way, while the participants were typically involved into other primary tasks. These comprised house work, surfing the internet, playing video games, and car driving. In general, this kind of situation occurs often and lasts long.

The role of music. The role of the music strongly depends on the respective situation. The participants' answers were mostly related to emotions. The most important effect of music was keeping, changing, or amplifying emotions. For example, in cases of parties, music played an important role in supporting good vibrations. When participants listen to music while relaxing, the music should calm them down. Another commonly named role was helping the participants in concentration while they were busy with another task, such as working or playing a video game.

Music players. Participants reported the use of a wide range of music players that fell into the classes of computer media player, portable mp3-player, CD-players, and radio. The participants mostly use pre-installed or old versions of computer media players. The vantages named by the participants about the system they use typically were associated with convenience aspects, such as simple user interface, familiarity, or immediately available music.

Handling patterns. There were certain patterns of using those music players identified. One pattern was that the participants did not really care about what music would be played. Therefore, they just select a radio station or a readily available playlist in order to listen to any music, immediately. The bigger part of the participants, however, picks a specific playlist, radio channel, or CD. Similarly often, the participants' generate playlists for certain cases, e.g. creating a playlist for chilling or a party. In these cases, the creating of the playlist was a central, artistic, and sometimes intimate process.

4.3.2 Comments on the prototype

Participants had the opportunities to express their desire for functionalities that they would like to see in a future version of the prototype. In addition we asked them to imagine potential advantages and disadvantages of a system similar to the presented prototype.

Desired functionality. Unsurprisingly the participants demanded the usual basic functionalities that are necessary to control music playback shown in Figure 3: Play, stop, pause, selecting the next or previous track, and change the volume. This function set is consistent with the function set reported by Kranz et al. [11] and Henze et al. [9].

Having an additional speech input to select a song by speaking the artists name and the album's title was named by three participants as well. Two participants wanted to be able to lay the CD on a table in order to support situational or congenital physical impairments. Two participants suggested extending the functionality to play other types of media, such as videos. Searching for related music was requested twice. Several other functionality was named only by a single participants, such as karaoke, displaying the title of the current song, using a mobile head-mounted web cam instead of a stationary web cam, show meta data about the current song, store the current playlist, toggle shuffle mode, alter the playlist, and having a timer.

Figure 3: The initial set of function that consists of: Play, stop, pause, next, previous, decrease volumn, and increase volumn.

Advantages and disadvantages. Beside the required functionality, we asked the participants to name the advantages and disadvantages they thought the prototype has compared to other solutions. The participants appreciated the very simple and visual interaction. They also appreciated the aesthetic aspect of holding the music in their hands. Furthermore, they expected it good for exploring their friends' music collections and that it would be more likely to listen to music they otherwise would not have listened to.

Most participants, however, had concerns that storing, managing, and handling the CDs would require lots of effort. One participant asked if she had to obtain the CDs by herself. Another concern was that it would be cumbersome for people that mostly listened to single songs that had no relation to full albums.

4.4 Discussion

Systems that demand repeatedly physical interaction, such as free-hand gestures, and handling a number of physical artefacts are not adequate for situations were music plays only a secondary role. In these situations participants pay only very little attention to the music and to controlling music playback. The played music must only roughly fit with the listener's taste and interaction with the playback device is often limited to turning the playback device on and off. When music plays a more important role users are willing to put more effort in controlling the music. We assume that a system similar to the used prototype could serve the user's need if the current handling pattern involves more complex but still simple tasks, such as selecting an album. A particularly promising application scenario for the intended system is, thus, a private party.

We observed that music listening is highly emotional and experiencing music not only with the ears but additionally by physically exploring and controlling the music might further support the emotional experience. If current handling patterns involves very intensive adjustment of the music playback we assume that users needs very fine grain control of the music. Limiting the functionality in any way seems to be not acceptable for the respective participants and a system using physical representatives for large collections of unstructured music no adequate solution.

The participants had the impression that the prototype is very easy to use. Nonetheless, some participants demanded complex and sophisticated functionalities. We observed a discrepancy between what the participants typically do with their music players and which functionalities they demand. If focusing on specific music listening situations the ease of use of the prototype can be retained by restricting the functionalities to a minimal set necessary in this context. It seems clear that users must be able to start, pause, and stop the music. In addition, selecting individual tracks from an album was a highly demanded feature that should be supported.

The participants' most serious concern was the handling and management of the CDs. Since the use of CDs to select albums is not necessarily a core aspect of the interaction design this can potentially be avoided. Another potential direction is the use of paper or plastic cards instead of audio storage media cover. These cards could be smaller than the necessary size of nowadays audio storage media cover. By reducing the artefacts size storing the artefacts would become less costly.

5. PARTICIPATORY DESIGN

The next step of the proposed process is to validate the initial set of functionalities and to derive multiple gesture sets from participatory design techniques. Accordingly, the aim of the second study was to derive free-hand gestures for the functions to control music playback found in the first study. Furthermore, the used function set, containing the functions play, pause, stop, changing the volume, and next and previous, is validated. A participatory design approach is used to collect gestures from potential users. The intended use case is to control music playback at private parties.

5.1 Methodology and Participants

Again the study was split into two parts. The goal of the first part was to validate the set of functions to control music playback. In addition, we collect further information about the music listening habits of the participants. To collect this information, semi-structured interviews based on a questionnaire are used.

In the second part we aimed at collecting potential gestures. Participants were asked to perform one gesture for each of the six functions. Thereby they "invented" gestures to control the music player. By using this participatory design approach we tried to ensure that intuitive gestures are derived.

In order to collect gestures for the intended situational context the study was conducted in the context of a private party but with one participant at a time. All participants were interviewed by the same person. During the first part the interviewer filled out the questionnaire and took additional notes, if needed. In the second part the participants were filmed by a webcam. 10 people from different educational and social backgrounds participated in this study. No participant has participated in the first study. Three participants were female and seven male. Their age ranged from 17 to 25 years with a mean of 20.8 (SD: 2.5). The study took about 10 to 20 minutes.

5.2 Apparatus

To provide feedback to the participants a wizard-of-oz approach is used. Participants performed the gestures in front of a monitor and equipped with a webcam (see Figure 4). A jewel case and a list of functions, that should be tried, were given to the respective person. The participants were asked to orally announce, which function they want to perform, and to use the jewel case, whenever they find it appropriate. The systems functionality was simulated by the investigator according to the respective function announced by the participant.

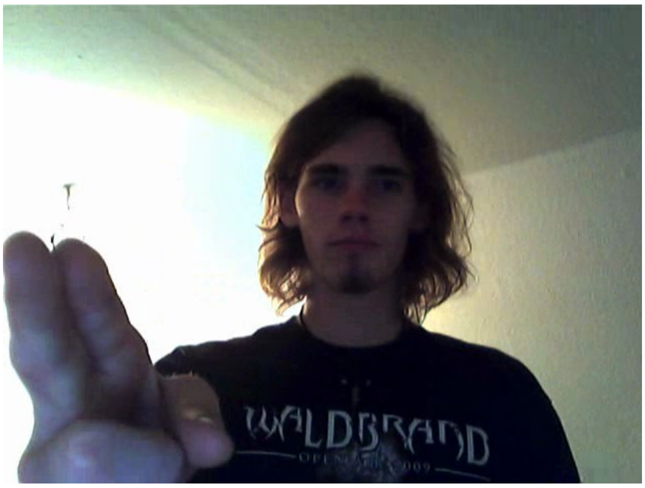

Figure 4: Exemplary camera image from the participatory design study.

5.3 Results

One person did not want to be filmed. Thus, only nine persons participated in the second part of the study. No significant differences were observed between groups formed by age or gender.

5.3.1 First Part: Interviews

The results for the function set are consistent with the first study. As the function set is also consistent with previous work [11, 9] we do not report the results in detail.

On a Likert-Scale from 1 (never) up to 5 (always) the participants rated how often they listened to music with M=4.2, SD=0.63. They listened to music from digital sources most of the time (M=3.6, SD=1.26), and less often from analogous sources (M=2.2, SD=1.23).

When asked which music player the participants use most often, with no answers given, six of the ten participants voted for the Windows Media Player. Three participants use Winamp and one iTunes.

5.3.2 Second Part: Inventing gestures

The jewel case was only used to switch to another album. For the other functions all participants performed gestures. For the analysis we differentiate between static gestures, dynamic gestures, and hybrid gestures. The three types are described in the following.

Used classification of gestures. Dynamic gestures are defined by the movement of the user's hands. The position of the hands and fingers is unimportant. An example is shown in Figure 5

Static gestures are the opposite of dynamic gestures. The posture of the hands and fingers is important, while the movement can be neglected. An example for this kind of

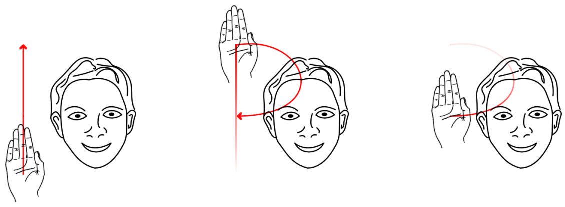

Figure 5: Example of a dynamic gesture

gestures is shown in Figure 6

Figure 6: Example of a static gesture

For hybrid gestures, both, the position and the movement of the user's hands, are important to recognize the gesture. The example in Figure 7 shows a right hand that performs a dynamic gesture, while the left hand performs a static gesture.

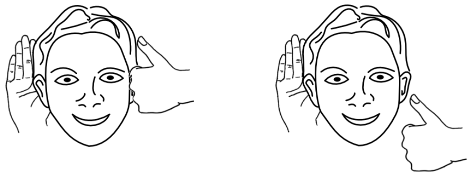

Figure 7: Example of a hybrid gesture

Play. One person covered the webcam with his hand to start the music. Two participants used static gestures. The first showed the *Victory* sign, the second built a triangle with his fingers, similar to the play symbol on a music player. Two persons formed one *pistol* or two *pistols* with one or both hands. These *pistols* first pointed to the ceiling and where then moved until they pointed at the webcam. Both variants are hybrid gestures. Another person who chose a hybrid gesture rotated his hands parallel 90° to the right. A fourth participant showed his fist and opened it to start the music. Dynamic gestures were performed twice. One of this gestures was again a *pistol* moved towards the webcam, while the other gesture was just moving both hands towards the webcam.

Pause. The same person that wanted to start the music by covering the webcam with his hand performs the same gesture to pause the playback. Three different static gestures were performed. The first participant formed the *Time-Out* symbol used in some sport games such as basketball. Another person formed a *pause* symbol by showing both hands with their sides to the webcam. The third person formed a very simple gesture by showing his fist to the webcam. Of the remaining five persons four performed a dynamic gesture by moving one or both hands towards the webcam. The fifth person moved both hands backwards.

Stop. Again the webcam was covered with a hand, this time to stop the playback. Static gestures were used twice. The first person formed a rectangle which should be similar to the stop symbol of a music player. The second person showed his fist to the webcam. Two gestures can be seen as hybrid gestures: One person hit his left hand with his right fist above his head. Another person acted like he was threatening somebody to cut his throat. The remaining four participants performed three different dynamic gestures to stop the music. The first person moved his hand from the right to the middle. The next person moved his hands towards the webcam. The last two people moved both hands like a *V*. Therefore they started with both hands in the middle of the body and then the right hand to the upper right and the left hand to the upper left.

Changing the volume. Using a static gesture, one person pointed with his index finger to the ceiling to increase the volume and pointed to the floor to decrease the volume. Three people performed hybrid gestures. The first participant turned up or down a fictive volume control to change the volume. Two other participants pointed with one thumb to the ceiling or the floor and shook the hand. The other five gestures were dynamic. One person moved his hands as if he was stretching something to turn up the volume and moved converse to decrease the volume. Another person seemed as if he was lifting or pushing down something with both hands to increase or decrease the volume. The last three participants moved their index fingers up to increase the volume. Two of them moved the hand down with the index finger pointing upwards to decrease the volume, while one person pointed down.

Next and previous. A static gesture was performed by pointing to the left or to the right with both hands to move to the next or previous track. The person who rotated his hands to start the playback this time uses only one hand for the same gesture to select the next track. A rotation by 90° to the left is used to select the previous track. The other seven participants performed dynamic gestures. The first person described clockwise circles with his hand to get the next track and counter-clockwise circles to get the previous one. Another person threw a fictive object over his shoulder left or right shoulder to go to the next or previous track. The remaining five people used some movement to the right or left to get the next or previous track. One of them started from the left or right and moved to the middle, while the others started in the centre. Three, including the one who moved from outer to centre, moved their hand with the side towards the webcam, one with his palm, and one pointed with his thumb into the direction he moved to.

5.4 Discussion

In the first part of the study we successfully validated the function set defined in Section 4. We assume that the reason why this set cannot be improved based on the results is that the participants of the first study already had a good understanding of the use case. Furthermore, this function set is also described in previous work and we therefore assume that it is the most basic function set needed to control music playback.

The fact that the jewel case was never used for another task but switching the album let us assume that the use of jewel case and maybe other objects to control the music playback might not be very intuitive. Many different gestures were used to do the same task, which was perhaps due to the fact that the participants were absolutely free to use any gesture. The gestures can be classified into three classes. Nevertheless no complete set of most used gestures was found, but some coincidences can be seen: Most dynamic gestures are kept very simple as movement along one axis. Also most static gestures are chosen as known symbols. In section 6 we will analyse the found gestures and deduce a set of gestures.

6. DEFINITION OF GESTURE SETS

The third step of the process is to formalize the proposed gestures and to define consistent sets of gestures. We found manifold gestures in the user study. Based on these gestures we define two consistent sets of gestures. To define consistent sets the first set consists of dynamic gestures only and the second set consist of static gestures. Most gestures were taken from the gestures proposed by the participants in the previous user study. Since we aimed at defining consistent gestures sets some gestures were chosen because they fit consistently with the other ones although the exact gesture was not proposed. In the following both sets are described.

6.1 Set of static gestures

Static gestures were not used as often as dynamic gestures. To form a consistent set, most gestures are similar to the symbols for the corresponding functions found on a music player.

Play. As shown in Figure 8.a a gesture was chosen that is similar to the symbol for *play* in various music players. There was no most used static gesture for this function, so this one was chosen, because it is one of the used gestures, easy to remember, and consistent with the gesture for *pause*.

Pause. Right to the gesture for *play*.b in Figure 8 is the static gesture for *pause*. As for the *start* function there is no most used gesture for this function. This gesture was chosen out of the used gestures in the explorative study, because it is easy to remember, simple to perform and consistent to the *play* gesture. Also this gesture is like a *weakened* version of the gesture for *stop*.

Stop. The static gesture for the stop function can be seen in Figure 8.c. This gesture is simple and known in other contexts. For example it is common to show both palms of the hands to somebody if you want him to stop.

Changing the volume. Most participants pointed up to increase the volume, not only when performing a clearly static gesture. For consistency the gesture for decreasing the volume is pointing down, as seen in Figure 8.d and Figure 8.e.

Next and previous. The only observed static gestures for these functions are shown in Figure 8.f and Figure 8.g. Again this is a gesture that is similar to the corresponding symbol on a music player. The two fingers showing in a direction are representing the two arrows of the next or previous symbol on a music player. As a result this gesture should be easy to remember.

6.2 Set of dynamic gestures.

Dynamic gestures were chosen more often in the user study. Most times the proposed gestures have one dominant direction. Although it is defined, that a dynamic function does not need a specific form how to hold the hand, one is given to ensure the comparability later in the evaluation.

Play and pause. The first column in Figure 9 shows how to start or pause the playback. Many participants did the same, or a similar gesture to start and to pause the playback, so these function are triggered with the same gesture. A movement towards the webcam was the most often seen gesture. We decided to use the most used form *pistol*.

Stop. In the second column of Figure 9 the gesture for the stop function can be seen. Most people used both hands for their *stop* gesture. The most used gesture is the *V* gesture, where both hands are moved from the centre middle to the outer left and right. A good alternative could be the static gesture combined with a movement towards the webcam.

Changing the volume. The most participants moved one or both hands upwards to increase the volume, respectively down to decrease. Additionally most people pointed up to increase or down to decrease the volume. The gesture which combines both, as presented in the third column of Figure 9, should therefore be very intuitive.

Next and previous. Almost all participants associated *next* with *right* and *previous* with *left*. Most of them moved their hand to the left or right. So this seems to be a strong connotation and should be used as dynamic gesture. The last column of Figure 9 shows that in this function the side of the hand is placed towards the webcam as this is the most used version of this gesture.

7. EVALUATION AND IMPROVEMENT

In the final step the defined gesture sets must be evaluated and eventually refined. Before implementing the gesture set defined in Section 6 they must be evaluated to test if one of them is more suitable than the other one. Furthermore, it should be determined if there are function besides the basic functions, that are essential for controlling a music player.

7.1 Methodology and Participants

The evaluation is split into three parts. In the first part the participants rated a set of given functions and had the opportunity to give new functions and rate them. In the second part both gesture sets were evaluated by each user. The participants were asked for missing functionality in the third part. After that the participants had the chance to give additional comments.

12 people with different backgrounds participated in this study. 5 of them were female and 7 male. Their age ranged from 17 to 25 years with a mean of 23.8 (SD: 3.8). The study took about 30 minutes. Even though, most of the people of the explorative user study announced, that they would like to participate in this evaluation, a divergent set of people was chosen for the evaluation.

7.2 Apparatus

The first part was a set of functions that could be rated using a Likert scale from 1 (unnecessary) to 7 (absolutely essential).

For the second part a *Wizard of Oz* prototype was built to simulate the functionality of both sets of gestures. This prototype was very similar to the one used in the explorative user study in Section 5 (see Figure 10). The set of functions differed from the one used at the study. Another difference is that during the evaluation an extra person was the *wizard* so that the interviewer could concentrate on the participants. Another rather obvious difference is that the participants this time had to perform given gestures. According to prin-

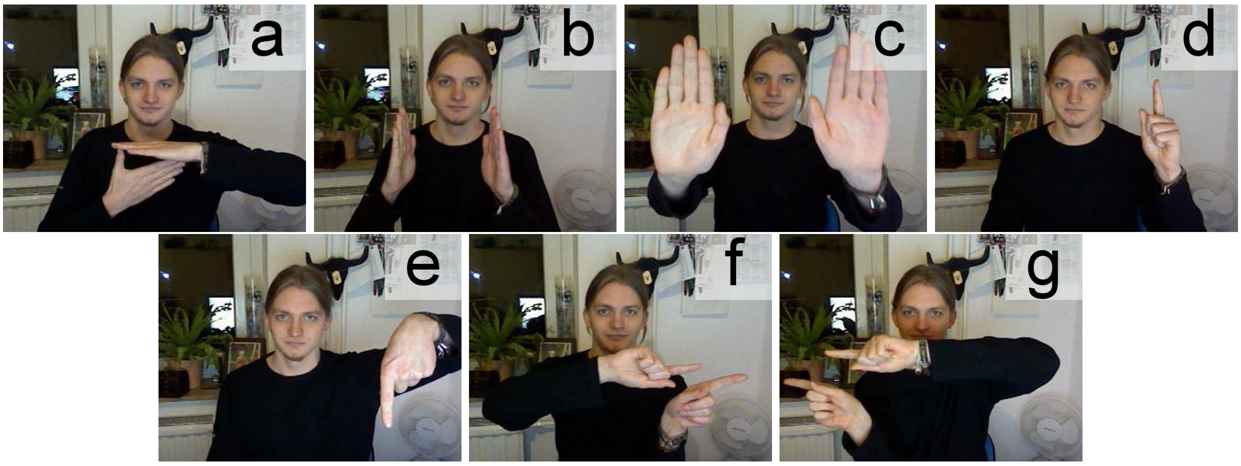

Figure 8: The set of static gestures (from a to g: Start, pause, stop, increase volume, decrease volume, next, and previous).

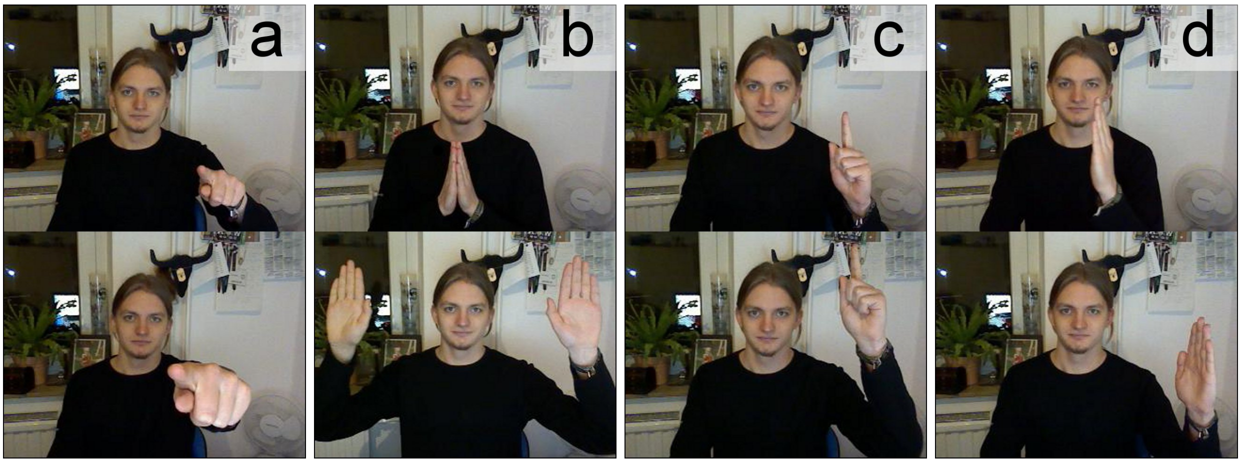

Figure 9: The set of dynamic gestures (from left to right: Start/pause, stop, increase volume, and next).

ciple of a *Within-Subject Design*, each participant evaluated both sets of gestures. To increase independency the order in which the gesture sets were evaluated switched after every trial. The participants were asked to perform each of the six defined gestures with both gesture sets. During the evaluation every person was recorded on video. After each trial the participants were asked to rate the gestures simplicity and intuitiveness. In addition, the consistency, rememberability, and delightfulness of the whole set was rated on a 7-point Likert scale (from 1 = not at all to 7 = perfect).

The third part was an interview on missing functions, where the interviewer took notes. If people had an idea for gestures that could be used to trigger a missing functions the according gesture was recorded on video.

7.3 Results

During the evaluation no significant differences were observed between groups formed by age, or gender.

In the first part our assumptions about the basic functions were confirmed. All basic functions were rated with a mean and median of at least 5.5, whilst amongst the other functions the maximum mean and median values were about 5.

Figure 11 shows the ratings for the two gesture sets. *Simple* and *Intuitive* refer to the means and medians of all functions of a gesture set. In the following the results are de-

Figure 10: Setup of the evaluation.

scribed in more detail.

Dynamic gestures. Most dynamic gestures were rated with a mean of 4 or above. Only the intuitiveness of the gesture for *stop*(M: 2.8, SD: 0.94) and the simplicity of decreasing the volume (M: 3.8, SD: 1.19) were rated less. On average the intuitiveness (M: 4.18, SD: 1.00) and simplicity (M: 4.52, SD: 0.78) of all functions were rated high. This set was also rated as consistent (M: 4.7, SD: 0.49). How easy the set is to be performed was not rated as good, but still not bad (M: 3.8, SD: 0.94). The joy of use seems to be high (M: 4.5, SD: 0.67).

Four times it was announced, that a good alternative for the stop function was to use a function similar to that one in the static set. Another participant mentioned, that he dislikes the gesture for *stop*. It was also suggested to spin a fictive volume regulator as an alternative to moving up or down to control the volume.

The taken videos show that most participants left the recorded area while performing a gesture at least once. Also the gestures were not performed very precisely. Sometimes the gestures for *next* and *previous* were not started from centre of their body. When performing the gestures for decreasing or increasing the volume, some participants shook their hands instead of moving them up or down.

Static gestures. Static gestures were on average rated less than dynamic gestures. Apart from *start*, all gestures were rated with a mean above 3.5. The intuitivity of *start* was not bad (M: 3.2, SD: 1.47), but the gesture was rated as the most difficult (M: 2.8, SD: 1.03). A very high rated gesture was *stop*. The gesture is the most intuitive (M: 4.8, SD: 0.39) of both gesture sets and the most simple one (M: 4.75, SD: 0.62) of the static gesture set.

It was announced, that the gesture for *start* was bad. Also few participants said that using one hand instead of both would have been as good, or even better. Five participants had problems to perform the gesture for *start*. Three test persons used their index fingers, instead of the side of their hands, to show the gesture for *pause*. Also the gestures often were not within the range of the video.

Dynamic vs. static gestures. Two-Tailed Dependent T-Tests were performed on the results. Though the gesture for *start* differs significantly with a $p < .01$, the coefficient of determination indicates with a value of $r^2 < .01$ that there is almost no linear correlation. On the first sight, this is surprising, because no participant rated higher for the static gesture, than for the dynamic one. This can be explained by the fact that all but three participants voted with a 5 for the dynamic gesture, while the ratings for the static gesture varied. We observed the opposite effect for the *stop* gesture. For this gesture the static one is significantly more intuitive ($p < .01$, $r^2 = .11$). Also the simplicity differs significantly, but again without a linear correlation ($p = .02$, $r^2 = .03$). The simplicities of the gestures for *next* and *back* have identical values. They differ significantly and show a weak linear correlation ($p < .01$, $r^2 = .28$).

Missing functions. In the last part the participant could give their opinion about functions that are missing so far. Two persons did not missed additional functions. The other participants provided many suggestions. For example most of the participants who would like to have a *random* function that can be toggled on or off. As gesture it was proposed to shake both hands as when dicing. A few persons proposed more complex functions, like editing the playlist. Therefore they would use a small set of gestures in the same way the few buttons on a MP3-Player are used to do that.

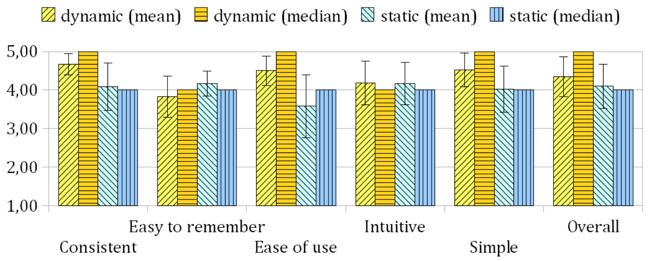

Figure 11: Concluded results of ratings for the gesture sets

Additional user feedback. Two participants said that a visual feedback was very important. Also two times it was said, that it would be best to mix both function sets. One person emphasised that it did not make sense to use gestures in any context. One person wanted to confirm the recognition of a gesture, before the system starts the function. The last feedback was that it was nicer to only use one hand for the gestures.

7.4 Discussion

Ratings. On average the dynamic set of gestures was rated higher, than the static gestures. But this result is not significant, as only a few gestures differ significantly, but not the whole sets. To make the dynamic gesture for *stop* more simple and intuitive, it could be tried to use the corresponding static gesture combined with a movement towards the webcam. For the gesture for *start*, a gesture could be used, that is inspired by the dynamic gesture. Maybe it is more intuitive and simple to point towards the webcam, than create a triangle with your hands. A revised set of gestures will be introduced in Section 7.5.

Missing functions. The creativity of the participants was surprising. Many functions and their gestures were suggested. Apart from simple function as e.g. repeating a track, or choose the next track randomly, some participants also suggested more complex functions, as seen in the results, that need menu-structures. It would be interesting to get to know, if the participants decided to use menus with gestures, because they are used to menu structures, or because it is intuitive.

7.5 Revised Design

To derive a final consistent gesture set, we combined the sets of gestures. Therefore we chose the gestures, that were observed to be the simplest and most intuitive ones for their function. After that the definition of some gestures were changed to be more clearly. To increase the chance of recognition for a system, we tried to create gestures, that have unique hand shapes, so that the movement is in most cases optional. The gestures introduced in the following might be less comfortable for a left-hander. That could be avoided, by also defining the gestures for *start* and *change volume* to work with the left hand.

Start. This gesture is defined as the dynamic one in Section 6. The difference is, that the movement is not needed. Therefore it is important to point with the right index finger towards the webcam.

Pause. This gesture is the same as the static one. It is similar to the gesture for *stop* so that they go well together. Also it is similar to the symbol for *pause* so that is it easy to remember.

Stop. Because the static gesture was rated much higher than the corresponding dynamic one, the static gesture is used for this function.

Next track. For this function the dynamic gesture is used. To the previous definition is added that the side of the hand must be the from the right hand, to make this gesture easier to recognize by the system.

Previous track. Analogous to *next track*. The movement is to the left and the left hand is used when performing the gesture.

Increase volume. When ignoring the movement, the dynamic and static gestures are the same for this function. To make it easier to recognise only the right hand should be used for this gesture.

Decrease volume. Analogous to *increase volume*, but this time the index finger of the right hands points to the bottom.

8. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a process to derive gestures from strong user involvements. The process is applied to the design of an interface to control music playback using free-hand gestures. Based on explorative user study the usage context is concretized and an initial set of necessary functions is collected. We successfully validated the function set in the subsequent user study and collected a large number of gestures for the functions using participatory design techniques. Two gesture sets are derived and evaluated in a comparative study. The results indicate that it is beneficial to derive more than one set of gestures to not exclude promising candidates for gestures.

With the conducted user studies we showed that the proposed process can be successfully used to define functionalities and a set of free-hand gestures. We assume that validating the design decisions in subsequent studies improves the final result. Throughout this paper, we intentionally refrain from considering technical limitations such as the performance of gesture recognition techniques. In future work it should be analyzed how the consideration of technical limitations from early on affects the outcome of the process.

Future work should also tell if the users' preferences for gestures change over time. In particular, we are interested in how the found gestures perform in long term studies. Furthermore, it should be investigated how a predefined set of gestures perform compared to gestures sets that are defined by individual users.

9. ACKNOWLEDGMENTS

This paper is partially supported by the European Commission within the projects InterMedia (FP6-038419) and HaptiMap (FP7-224675).

10. REFERENCES

[1] D. Akers. Wizard of oz for participatory design: Inventing a gestural interface for 3d selection of neural pathway estimates. In *Ext. Abstracts CHI*, 2006.

[2] T. Baudel and M. Beaudouin-Lafon. Charade: remote control of objects using free-hand gestures. *Communications of the ACM*, 36(7):35, 1993.

[3] R. Bolt. "Put-that-there": Voice and gesture at the graphics interface. In *Proc. SIGGRAPH*, 1980.

[4] EyeSight. Symbian moove - gesture controlled music player, http://www.eyesight-tech.com, 2010.

[5] W. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In *Proc. FG*, 1995.

[6] W. Freeman and C. Weissman. Television control by hand gestures. In *Proc. FG*, 1995.

[7] K. Hayafuchi and K. Suzuki. MusicGlove: A Wearable Musical Controller for Massive Media Library. *Pro. NIME*, 2008.

[8] N. Henze and S. Boll. Snap and share your photobooks. In *Proc. ACMMM*, 2008.

[9] N. Henze and S. Boll. Designing a CD augmentation for mobile phones. In *Ext. Abstracts CHI*, 2010.

[10] T. Hesselmann, S. Flöring, and M. Schmitt. Stacked half-pie menus: navigating nested menus on interactive tabletops. In *Proc. ITS*, 2009.

[11] M. Kranz, S. Freund, P. Holleis, A. Schmidt, and H. Arndt. Developing gestural input. In *Proc. IWSAWC*, 2006.

[12] C. Kray, D. Nesbitt, J. Dawson, and M. Rohs. User-defined gestures for connecting mobile phones, public displays, and tabletops. In *Proc. MobileHCI*, 2010.

[13] R. Liang and M. Ouhyoung. A Real-Time Continuous Gesture Recognition System for Sign Language. *Proc. FG*, 1998.

[14] T. Masui, K. Tsukada, and I. Siio. MouseField: A Simple and Versatile Input Device for Ubiquitous Computing. *Proc. UbiComp*, 2006.

[15] T. B. Moeslund, M. Störring, and E. Granum. A natural interface to a virtual environment through computer vision-estimated pointing gestures. *Gesture and Sign Language in Human-Computer Interaction*, pages 59–63, 2002.

[16] M. Morris, J. Wobbrock, and A. Wilson. Understanding Users' Preferences for Surface Gestures. In *Proc. GI*, 2010.

[17] M. Nielsen, M. Störring, T. Moeslund, and E. Granum. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In *Gesture-Based Communication in Human-Computer Interaction*, pages 105–106, 2004.

[18] J. Rico and S. Brewster. Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability. In *Proc. CHI*, 2010.

[19] M.-C. Roh, S.-J. Huh, and S.-W. Lee. A virtual mouse interface based on two-layered bayesian network. In *Proc. WACV*, 2009.

[20] T. Schlömer, B. Poppinga, N. Henze, and S. Boll. Gesture recognition with a wii controller. In *Proceedings of TEI*, 2008.

[21] B. Stenger, T. Woodley, and R. Cipolla. A Vision-Based Remote Control. *Computer Vision*, pages 233–262, 2010.

[22] D. Wang. Giuc: A gesture interface for ubiquitous computing. In *Proc. CMC*, 2009.

[23] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proc. CHI*, 2009.