Gesture-Aware Remote Controls: Guidelines and Interaction Techniques

[1]Gilles Bailly [2]Dong-Bach Vo [2]Eric Lecolinet [2]Yves Guiard
[1] Deutsche Telekom Laboratories, TU Berlin, Ernst-Reuter-Platz 7, D-10587, Berlin, Germany
2 Telecom-ParisTech, 46 rue Barrault, 75013 Paris, France
Gilles.Bailly@ telekom.de {Dong-bach.Vo, Eric.Lecolinet, Yves.Guiard}@telecom-paristech.fr

ABSTRACT
Interaction with TV sets, set-top boxes or media centers strongly differs from interaction with personal computers: not only does a typical remote control suffer strong form factor limitations but the user may well be slouching in a sofa. In the face of more and more data, features, and services made available on interactive televisions, we propose to exploit the new capabilities provided by gesture-aware remote controls. We report the data of three user studies that suggest some guidelines for the design of a gestural vocabulary and we propose five novel interaction techniques. Study 1 reports that users spontaneously perform pitch and yaw gestures as the first modality when interacting with a remote control. Study 2 indicates that users can accurately select up to 5 items with eyes-free roll gestures. Capitalizing on our findings, we designed five interaction techniques that use either device motion, or button-based interaction, or both. They all favor the transition from novice to expert usage for selecting favorites. Study 3 experimentally compares these techniques. It reveals that motion of the device in 3D space, associated with finger presses at the surface of the device, is achievable, fast and accurate. Finally, we discuss the integration of these techniques into a coherent multimedia system.

Categories and Subject Descriptors
H5.2. **[Information interfaces and presentation]**: User Interfaces-Evaluation/methodology, Interaction styles.

General Terms
Human Factors

Keywords
Mid-air gestures, remote control, 10-foot interaction, menu, ITV.

1. INTRODUCTION

Interaction with TV sets, set-top boxes or media centers strongly differs from interaction with personal computers. This context of use, which we believe has received too limited attention so far in HCI, is quite challenging [3] [6]. The number of available features and services has been uninterruptedly increasing. Not only hundreds of channels are now available through cable TV but also Web services provide various kinds of digital media (podcasts, video on demand, digital stores...). Moreover, set-top-boxes can also provide access to personal data such as photos, music, etc. The WIMP model, that relies on efficient devices such as a mouse and a keyboard makes it reasonably easy to interact with numerous sources of data on the PC. However, this model is not adapted to the context of TV and "sofa interaction" [3][6]. They must usually deal with much more limited devices, such as remote controls and must thus navigate long lists or deep hierarchical menus by making repetitive presses on a directional pad or to cope with remote controls that are overcrowded with buttons. Moreover, users can be leaning back on a sofa. In either case we face a usability problem [8][15][37]. This is especially true for favorite commands that users perform frequently (e.g. for displaying their favorite TV channels or accessing their favorite Web services, personal data, etc.).

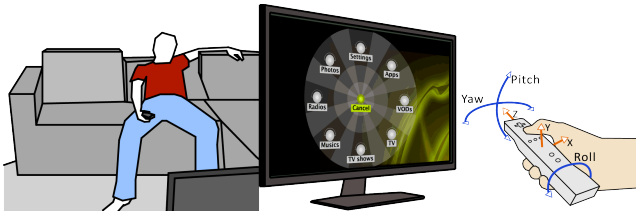

Fig. 1. Left: Sofa interaction. Right: Augmenting a remote control with mid-air gestures allowing 6 degrees of freedom.

One solution for augmenting the interaction bandwidth consists in transforming the traditional remote control into a sophisticated device including a touchscreen [22] or even in using a smartphone or a tablet as a replacement [21][38]. We focus on another approach that exploits the new capabilities provided by gesture-aware remote controls [13][33]. A few prototypes and even some commercial products [11][16] have been recently introduced, but to date little HCI research has been done on interaction techniques with remote-controls [3] and gesture-augmented devices.

Mid-air gestures are known to be well accepted by users in the domestic context [14][30] and they offer several notable advantages. Unlike buttons and other physical interactors, they consume no real estate on the remote control and so the input vocabulary can be enriched at no cost with regard to device size. They do not confiscate visual attention—the users' gaze need not alternate between the TV screen, their only source of interest, and the remote control. Previous research on 2D gestures, especially on Marking menus [1][20], suggest that certain gestures are easy to remember, a property that makes them good candidates for acting as shortcuts for performing favorite commands. Finally, mid-air gestures only require low-cost sensors such as accelerometers or gyroscopes.

We thus investigated how mid-air gestures can augment interaction with remote controls in the context of interactive television (Fig. 1). Study 1 shows that the optimal mapping between mid-air gestures and directional actions depends on context and that pitch & yaw rotations should be preferred for remote control interaction. We then focused on eyes-free roll

gestures. Study 2 indicates that users can accurately select up to 5 items, up to 7 items if the system takes into account the identity of the user. The data we gathered helped us to design a vocabulary of simple mid-air gestures combinations. We proposed five different techniques using either device motion, button-based interaction, or both. Study 3 experimentally compares them. It reveals that motion of the device in 3D space, associated with finger presses at the surface of the device, is achievable, fast and accurate. Finally, these techniques were implemented in the context of interactive television for navigating in multimedia data and quickly selecting favorite elements. They make gestures visible and easy discoverable [25] and favor the fluid transition from novice to expert usage [18].

2. RELATED WORK

2D or 3D gestures can be conveniently classified as symbolic, physical, metaphorical or abstract [37]. An example of *symbolic* gestures is provided by users who enter numbers by drawing their shape in the air using the MagicWand [7]. *Physical* gestures work as their counterpart in the real world as in [2], where users manipulate virtual 3D objects. *Metaphorical gestures* transpose physical gestures in the digital world, as the swipe gesture in 3D space to go to the next chapter as if browsing a real book [4] or to go to the next slide with the Gyration mouse [11]. For these three gesture categories, however, there is little coherence among users [14]. With *abstract gestures*, on which we will focus below, the mapping is arbitrary between the gesture and its semantics. A well-known 2D-space instance is the Marking menu [18] which maps gestures onto an arbitrary arrangement of actions. Although proved to be remarkably efficient [1] [18] [39], the Marking menu has received limited attention in the field of 3D interaction where techniques generally rely on the other three categories.

Mid-air gestures on mobile devices. Motion Marking menus [27] are an adaptation of Marking menus for mobile devices where users perform pitch rotations to select commands in a 3x3 hierarchical menu system. Mid-air gestures have also been used to enter text on mobile devices [28] [31] [34] [35] Unigesture [31] uses pitch and roll rotations for choosing a group of characters and an inference engine to overcome ambiguities. TiltType [28], TiltText [35], and Shrimp [34] combine button clicks for choosing a group of characters and pitch and roll rotations for further choosing the desired character within the selected group.

Mid-air gestures on remote controls. Text entry using a gesture-sensitive remote control has also been investigated. In GesText [13] users combine pitch and roll gestures, without the help of any buttons. With Cube Key [32], each letter (placed in a 3x3x3 matrix) can be accessed from translations in 3D space. Gesture-sensitive remote controls are well known in the video game industry, following the example of Nintendo Wii remote control (Wiimote). XWand [36], which works in an intelligent environment, is used to select elements by pointing and to execute a small set of commands using roll gestures. A likely reason why most of the above-described techniques use pitch and roll rotations is that they rely on accelerometers, which cannot detect yaw rotations. We will see below that this technological constraint may appreciably impact usability in the case of remote-control interaction.

Vision-based remote controls. Various vision-based solutions, exploiting a video camera, have been proposed for remote control systems [5] [10] [16], most relying on hand-gesture interaction. These techniques do not provide tactile feedback and they can introduce mode-switching problems due to the lack of buttons as delimiters. Typically, they offer a limited set of commands, although the domain is evolving at a fast pace [10] [16]. Besides, if our study considers interaction with a hand-held device, our results are relevant to gestural interaction in general, our focus being the set of gestures rather than the hardware technology.

3. STUDY 1: MAPPING PREFERENCES FOR MID-AIR MOTION

Six geometrical degrees of freedom (DoF) are available for the motion of a remote control (Fig. 1), but the question whether it is possible to base interaction on a distinction between rotations and translations is open, and the sort of motion users actually produce may be context dependent (e.g. when they use a remote control vs. a smartphone). This exploratory investigation considered an eight-choice selection to be made by means of gestures. Items appeared at 8 possible locations on a 3x3 matrix whose central cell was empty. With such an arrangement each item could be identified just as well by its cartesian (*x*-*y*) or polar (directional) coordinates. So the task, selecting items in menu systems, did not bias participants toward performing translations (x, y or z) or rotations (pitch, yaw, roll).

Methods. For each trial, one matrix cell was highlighted at chance. Following the approach proposed in [37], "first portray the effect of a gesture, and then ask users to perform its cause", we asked our participants to perform the mid-air gesture that, in their opinion, should have caused this effect. We also explained them that the central cell corresponded to the resting position (i.e., when the device was roughly parallel to the floor) and that they had to start from this position to perform the gesture. They had to come back to the resting position before performing the next trial. The sessions were videotaped with participant's informed consent.

While previous studies dealing with this sort of tasks were run with handheld devices, we tested our participants with a smartphone (an iPhone, S condition) vs. a remote control (a basic model with no screen, RC condition). In the S condition, the stimuli appeared on the device they were manipulating. In the RC condition the stimuli appeared on a 3m-distant (10 feet) TV screen and so the participants never had to look at the remote control.

We used a between-participant design. For each condition (S and RC), we recruited 9 participants (18 in total) aged 24-30. Each participant performed 8 gestures, one for selecting each matrix cell.

Gestures were classified from videos in one of the 6 following categories: X, Y, Z translations and yaw, pitch, roll rotations. This classification was performed separately by two different people and the same results were obtained in both cases. As this classification was done from naked-eye observations it only accounts for the main tendencies (visually indiscernable small auxiliary movements may have also been produced).

Results. While 7 of the 9 participants of the Smartphone group demonstrated conspicuous pitch and roll rotations, the pattern was opposite in the Remote-control group, in which 7 of the 9 participants showed no less conspicuous pitch and yaw rotations ($p<.001$ by the Fisher exact probability test). Two participants showed translations in each of the two groups.

Implications for HCI. This exploratory study delivered three suggestions of potential relevance to the design of a mid-air gesture vocabulary:

(1) The participants made gestures that were unambiguously of either the rotational or translational category. The behaviors were consistent, no participant switched from one category to the other during the experiment.

(2) Fourteen or our 18 participants spontaneously demonstrated rotations and only four demonstrated translations ($p<.04$ by the

binomial probability test, two-tailed). This outcome, perhaps simply a reflection of the fact that rotations are easier to perform for a multi-joint arm [12], encouraged us to investigate rotations in the rest of the present study.

(3) Different sets of rotations were performed by participants in the S vs. RC conditions: while all the rotations observed in the S group combined the pitch and the roll (in keeping with previous interaction techniques on smartphones [28][31][35]), all those observed in the RC group combined the pitch and the yaw. If this result suggests that accelerometers, which capture pitch&roll, are sufficient for smartphones, it points to the utility of gyroscopes, which can also detect yaw gestures, for directional tasks with remote controls.

None of our participants demonstrated translations along the z axis or roll rotation in the RC condition. This means that these two components remain free for performing other tasks while combined with pitch and yaw movement. As z-translations require the combined motion of both the elbow and shoulder joints, they are likely to be cumbersome for seated users. This is why we decided to focus on roll gestures in the rest of the study.

4. STUDY 2: USING ROLL GESTURES

Device rotations about the roll axis (Fig. 1), which results from wrist pronation and supination, have been considered in a number of papers. While Rahman et. al.'s users were able to select 16 items by rolling a smartphone with visual feedback [29], only three levels of roll movements were used in [26], which proposed an eyes-free menu technique. This rather large difference is easily explained by the absence of visual feedback in the latter case, but we suspected that users should be able to select a larger number of commands eyes-free with a reasonable level of accuracy. This led us to design two user studies to explore the potential of roll gestures for eyes-free selection. We used a TV screen to deliver stimuli, our participants being asked to perform the required roll movements with a Wiimote.

4.1 Study 2A

This user study aimed to evaluate the impact of two factors, *number of items* and *user posture*, on the feasibility of eyes-free roll selection.

Number of items. We considered four different set sizes with 5, 7, 9, and 11 items, leaving aside the easy three-item condition tested in [26]. We expected our participants to be able to accurately discriminate up to 7 or 9 different eyes-free roll movements.

Posture. Roll movements may imply several joints, the wrist, the elbow and even possibly the shoulder for very large amplitudes. In the context of 'sofa interaction' with a remote control, motion about these joints must be more or less constrained by the user's posture. We asked participants to perform the rolls in three different postural conditions:

- Posture 1: only the wrist is free, the forearm resting on the thigh.
- Posture 2: the upper arm is kept in contact with the body so that only the elbow and the wrist can freely move.
- Posture 3: all three joints are free to move.

Roll movement. We used a *press-tilt-release* strategy for measuring the roll movement [29]: the captured value was the amount of angular change recorded with the Wiimote from an initial angular position marked by a trigger press to a final angular position marked by trigger release (Figure 2-right). This strategy has two main advantages. One is that the input value is relative, meaning one need not care about the absolute vertical, unlikely to be correctly perceived by slouching users. The other is a considerable increase of the range of available angular values. For example, to produce a large-amplitude roll in the clockwise direction the gesture begins with a preparatory counter-clockwise roll at the end of which the trigger is pressed, setting an angular zero, say, at -45° (Figure 2-left); if, the trigger being kept pressed, the device is then rolled clockwise up to +75° and the trigger is released at that new position, the obtained angular delta will be 120°. With this technique the possible range can then reach 200° and even more (depending on the user and the posture).

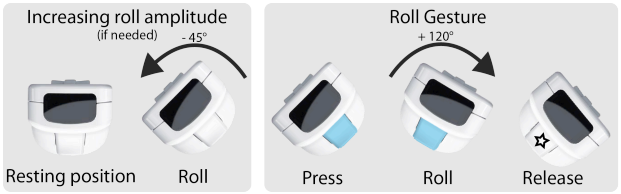

Figure 2: Example of a roll gesture. Left: a preparatory back-roll for increasing the total roll movement. Right: Execution of the roll gesture (the trigger being kept pressed).

Stimulus and Feedback. We used a TV screen to display a horizontal array of constant-sized boxes. A variable number of boxes served to vary the number of items that had to be discriminated using roll movements. An example with seven items (7 boxes) is shown in Fig. 3. The stimulus consisted of one square being highlighted. If the central square was highlighted, no roll was needed, the participant having just to press and release the trigger. The position of the highlighted box served to specify the direction and the relative amplitude of the required roll movement without providing any cue about where in the absolute range of positions the angular variation was to be produced. In practice, the more to the right (resp. left) the stimulus, the larger the roll amplitude the participants had to cover in the clockwise (resp. counter-clockwise) direction from their initial trigger press to their final trigger release. Apart from proprioception no feedback was provided in this experiment. Indeed, we wished to determine which angular changes users would spontaneously perform, and which patterns the distribution would exhibit.

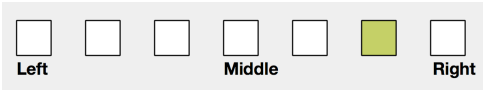

Fig. 3. The stimulus for 7-item selection. The highlighted box indicated which item of the set had to be selected (left = counter-clockwise, right = clockwise).

Task. In response to each new stimulus the participant could prepare his/her gesture by changing the starting position if necessary and then perform a *press-tilt-release* roll (Figure 2) as fast and accurately as possible. An inter-trial interval of 1.5 seconds was used to give time to the participant to return to their rest position before the next trial.

Participants and Apparatus. Nine volunteers (aged 22-28, 8 men) were recruited from our institution. The experiment was performed using a MacBook Pro and a 40" LCD HD TV set. Participants were seated in a couch, about 3m (10 feet) away from the TV set. The remote control was a Nintendo Wiimote. The experimental software (C++/Qt) used an adaptation of the DarwiinRemote framework [9] to communicate with the Wiimote.

Procedure. The experimental session started with a few warm-up trials using a three-item set. Each participant completed 5 blocks for each combination of item set size and posture. Within a block

each item appeared once in a randomized succession. The four set sizes (5, 7, 9, and 11 items) were presented in increasing order, task difficulty thus increasing with familiarity. The posture was counterbalanced across participants using a Latin Square design. In summary, the experiment involved:

9 participants x 4 item-set sizes x 3 postures x 5 blocks x 5, 7, 9 or 11 items (depending on item-set size) = 4320 gestures. The total duration of the experiment was about 45min.

4.1.1 Results

Angular ranges. The participants having received no indication about the angular amplitudes they had to cover with their rolls, considerable differences were observed among them. For instance, the angular max-min range varied from 162° to 326° in the 7-items condition (the max-min range could reach 326° because we used a *press-tilt-release* strategy, as explained above). We normalized the angular variation by dividing it by the observed max-min range for each participant and each item-set size. Unsurprisingly, the max-min angular range was strongly dependent on item-set size ($F_{3,24}$=97.47, p<.0001) and increases with the *number of items* (5: 159°, 7: 306°, 9: 343°, 11: 371°). The max-min angular range depended slightly but significantly on *posture* ($F_{3,24}$=9.8, p<.001), with on average 273° for posture 1, 290° for posture 2, and 321° for posture 3.

Table 1: mean angular variations (in degrees) for each item.

Nb of items	Item identity										
	-5	-4	-3	-2	-1	0	1	2	3	4	5
5				-81.2	-43.6	-0.3	42.8	78			
7			-155	-90.3	-49.6	-0.9	52.9	88.6	150.2		
9		-173	-116	-83.1	-51.3	-0.6	50.9	81.9	118.5	171	
11	-188	-136	-101	-77.5	-46.8	-0.6	46.4	70.6	99.3	138.8	184.4

Angular Variations. Table 1 shows the average angular variation for each item and for each item set size. Participants adjusted the angular amplitude of their rolls proportionally to the number of items. A nearly perfect linear relationship was found for each of the four item-set sizes (all r^2>.995), with a virtually zero *y*-intercept (all four intercepts being less than 1° and non-significant). Moreover, the angular change between two contiguous items was fairly constant, about 45° (especially in the 5 and 7 item conditions).

Recognition rates. We used a KNN algorithm [17] with *k*=5 and the Euclidian distance on the angular variation to calculate the recognition rate. KNN is useful to reduce the effect of noise on the classification [17] and is fast and effective [1]. The testing and learning bases were separated by using a leave-one-out cross-validation technique [23]. There was a strong effect of item-set size on accuracy ($F_{3,24}$=53.52, p<.0001). Accuracy decreases with the *number of items* (5: 96.3%; 7: 87.7%; 9: 77.1%; 11: 63.1%). No significant effect of posture was detected.

4.1.2 Implications for HCI

This user study led to four informative findings.

(1) Whether or not users are provided with feedback impacts the relationship between angular variation and item-set size. While Rahman [29] found a *quadratic* relationship with a continuous visual feedback, our results reveal a *linear* relationship when gestures are performed eyes-free.

(2) The angular change between two contiguous items was fairly constant, about 45° (especially in the 5 and 7 item conditions).

(3) Surprisingly, the posture did not impact the recognition rate while it had an effect on maximal angular variation. As noted, participants divided the available amplitude in about equal sectors, whose size depended on posture. This is encouraging as it suggests good robustness.

(4) Users were able to easily select five items by performing eyes-free roll gestures with recognition rates above 95%. The conditions with 9 and 11 items turned out to be quite difficult without visual feedback. The seven-item case (87.7%) seemed inconclusive. This led us to the next study.

4.2 STUDY 2B: FOCUS ON 7-ITEM

Participants were now presented with numeric values on the TV screen (-135°, -90°, -45°, 0°, 45°, 90°, 135°). These values were chosen in light of the results of the previous experiment, the step values being simply rounded up to 45° to facilitate mental representation. We no longer inquired into the posture issue.

A sample of 20 fresh volunteers (aged 25-31 years, all right-handed) participated. The experiment comprised a learning and a testing phase. During the 5min learning phase, the value of the angular variation actually covered by the participant was displayed on the screen (this feedback was no longer delivered during the testing phase). Each participant ran 15 blocks in each of which the 7 items appeared once in a randomized order. In summary, the experiment involved 20 participants x 15 blocks x 7 items = 2,100 selections.

4.2.1 Results

Accuracy. As previously, we used the KNN algorithm with *k*=5 and a leave-one-out cross-validation technique [23]. Recognition rate ranged from 88 to 95%, with an average of 90.9%. We also analyzed the data by participant, that option corresponding to the case where the recognizer already has information about user identity (e.g., if a user profile has been selected or thanks to an automatic face recognition system). We then obtained an accuracy rate of 96% (from 91.4 to 98.1%).

Selection Time. Measured from stimulus appearance to button release, selection time was, as expected, strongly dependent on item position ($F_{6,114}$=32.4, p<.0001). On average, the central item (1.40s) was significantly faster than items -1 (2.44s) and +1 (2.60s), both significantly faster than items -3 (3.53s) and +3 (3.27s). The central item was also significantly faster than items -2 (2.67s) and +2 (2.98s).

Angular variation. Table 2 shows the mean angular variations, which will serve to parameterize the roll-based techniques.

Table 2. Mean, 5th percentile, and 95th percentile of angular change (in degrees) for each item

Item	-3	-2	-1	0	1	2	3
delta angle (mean)	-134	-85	-51	0	49	86	127
delta angle (p05)	-164	-106	-70	-4	31	61	100
delta angle (p95)	-109	-66	-32	4	70	111	166

4.2.2 Summary

Our investigation of eyes-free roll selections revealed that this approach is fairly robust, as the posture does not impact the recognition rate. It turned out that all users could easily perform five-item selections and even seven-item selections if the database is adapted to the user (i.e the system takes into account the identify of the user). Seven items appeared to be an upper limit for eyes-free roll gestures.

5. DESIGN SPACE & TECHNIQUES

Building on the results obtained in previous sections, we now propose several multimodal techniques for interacting with a remote control. All of them aim at selecting items in a two-level hierarchical menu and can work both in novice mode (with the menu shown on the TV) and in expert mode (eyes-free). We will describe some novice-mode visual representations, but let us start with eyes-free selection. Even though visual feedback is needed for some commands in a complex interactive multimedia system, open-loop selection without the burden of TV menu navigation is especially desirable for frequent actions like selecting a *favorite* channel.

Table 3: Design space for a two-level menu. Techniques that appear over a white background are evaluated in the next section.

2nd. level ⇒ ⇓ 1st. level	*Pitch & Yaw*	*Roll*	*Pad*
Pitch & Yaw	P&Y/P&Y	P&Y/Roll	P&Y/Pad
Roll	Roll/P&Y	Roll/Roll	Roll/Pad
Pad	Pad/P&Y	Pad/Roll	Pad/Pad

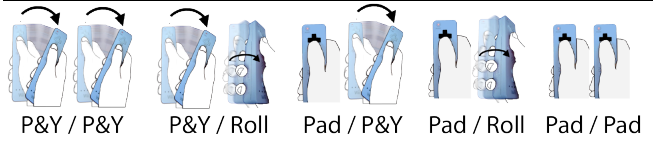

Figure 4: The gestures of the five two-level menus

The foregoing suggests that three gestural modalities, pitch&yaw, rolls, and button interaction, are of special interest. Table 3 combines them. As each menu level can be controlled by one of three modalities, nine combinations are theoretically possible. Below we describe each modality and then explore the design space.

Pitch and yaw gestures *(P&Y)* involve two rotational DoF. Yet we will consider them here as a single input modality because they combine in an integral way for the task under consideration (a 2D directional task performed relatively to the TV screen). Any combination of pitch and yaw rotations specifies a movement in a certain direction as in Marking menus [18]. Marking menus have been shown to be very efficient in 2D space, especially for eyes-free interaction, and so expanding them to mid-air interaction looked especially promising. Note incidentally that our transposition of Marking menus to 3D space is quite different from Motion Marking menus [27] and closer to the original.

Roll gestures (*Rolls*) correspond to the second modality and will be used as explained in the previous section.

The third modality, Pad gestures *(Pad),* consists in pressing the buttons of a directional pad (d-pad). D-pads are quite common in remote controls and are well suited for directional tasks. This latter property is interesting because it will allow direct comparison with pitch and yaw gestures, thus making it possible to answer the question whether gestural interaction is competitive compared with button-based interfaces. Moreover, pad gestures can easily be combined with gestural modalities.

5.1 Interaction Techniques for 2-Level Menus

P&Y / P&Y. This technique is reminiscent of Multi-Stroke menus [39] in the sense that user makes two successive gestures, but these are pitch and yaw gestures. Each gesture, as already described, consists of pressing the trigger button at the beginning of the gesture and releasing it at the end.

P&Y / Roll. Different gestures are then used for the first (P&Y) and second (Roll) level of the menu system. A theoretical advantage is that these gestures do not need to be performed sequentially as they imply different modalities.

Roll / P&Y. This technique was discarded because pretests showed the difficulty of performing roll *then* pitch&yaw gestures. The case of performing roll and pitch&yaw gestures simultaneously is already covered by the previous technique.

Roll / Roll. This technique was discarded based on poor pre-test results. Both steps require angular variations that theoretically do no depend on each other, but many participants were confused, failing to perform the second gesture. Often, they felt the need to perform the second gesture according to an absolute vertical position, thus wasting time. And the users that did not follow this pattern made many mistakes.

Pad / Pad. This technique acts as a baseline and we expected it to be especially efficient, as the user just needs to click two times on the appropriate buttons of the d-pad.

Pad / P&Y. The promise of this technique is that it combines d-pad and mid-air gestures. It also saves one action because the d-pad also serves as an on/off control.

Pad / Roll. This technique is similar to *Pad / P&Y* except that the second gesture is a roll.

P&Y / Pad. Although this technique looks similar to *Pad/P&Y*, it worked poorly in our tests and we discarded it. In Pad/P&Y, the d-pad press and release actions serve as start- and end-delimiters for the mid-air gesture. But P&Y/Pad does not allow that. Another button (such as the trigger) is then needed to delimitate the gesture (this being needed to improve the precision and avoid accidental detections whenever the device is moved). Asked to press several buttons in the correct order while performing mid-air gestures, participants made too many errors for this technique to deserve further consideration.

Roll / Pad was discarded for the same reason as *P&Y / Pad*.

In sum, we retained the *five* interaction techniques of Figure 4, combining the 3 above-described modalities: *P&Y/P&Y*, *P&Y/Roll*, *Pad/Pad*, *Pad/P&Y and Pad/Roll*.

5.2 Experiment 3

The goal of this experiment was to compare the efficiency of the five above-described techniques in expert, eyes-free mode. As noted above, the Pad/Pad technique may be seen as a nearly ideal case serving as a baseline.

Participants and Apparatus. 13 adult volunteers (all right-handed, seven male) participated in a single 45mn session. The same equipment and software was used as in the previous experiment except that to capture yaw gestures we used the MotionPlus expansion of the Wiimote, which carries a gyroscope.

Number of items. Because the Wiimote d-pad, like most common d-pads, has only four directions, all menus were limited to 4 items. Cardinal directions on the main axes were used for the *P&Y* modality, and 90° and 45° items on each side of the neutral position for the *Roll* modality. All menus thus contained 4x4 = 16 items.

Task. We used the symbolic stimuli in Fig. 5. For each trial, two symbols were displayed on the center of the screen. The first symbol corresponded to the first modality, the second symbol to the second modality (for instance, we used the stimuli Fig. 5-a and Fig. 5-b for the Pad-Roll technique). We used different symbol for

each modality to increase the mapping between symbol and the required gesture and to avoid possible biases. The participant then performed the corresponding gesture. The gestures actually recognized by the system were then displayed on the TV screen for 3 seconds. This visual feedback was provided in order to help the user to improve the gestures and to better represent real usage, as some sort of a feedback is always obtained in real conditions. The 3s delay also ensured that the participant had time to come back to the rest position.

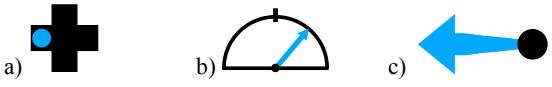

Fig. 5: Stimuli for: a) pad; b) roll; c) pitch&yaw modalities.

Design. Following instructions, participants were offered the possibility of performing up to three warm-up blocks per technique. The experiment was organized in five blocks. For each block, the 16 items appeared one time in randomized order. The order of technique was counter-balanced across participants using a Latin Square design. In summary, the experiment involved:

13 users x 5 techniques x 5 blocks x 16 items = 5200 selections.

At the end of the experiment, participants were given a questionnaire to investigate subjective preferences, which covers speed, complexity and fun about the technique.

5.3 Results

Speed (Fig.6). An ANOVA revealed a significant effect of technique ($F_{4,48}$= 25.34, p<.0001). Post-hoc Tukey tests showed that Pad/P&Y (1.39s) and Pad/Pad (1.40s) were significantly faster than Pad/Roll (1.65s) and P&Y/P&Y (1.68s), both significantly faster than P&Y/Roll (2.02s).

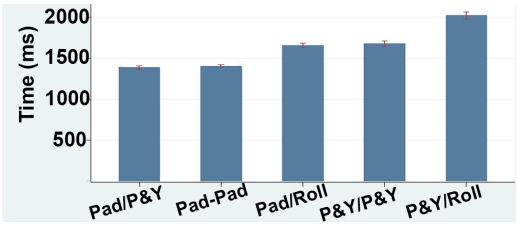

Fig. 6: Mean selection time for each technique with 95% confidence interval marked.

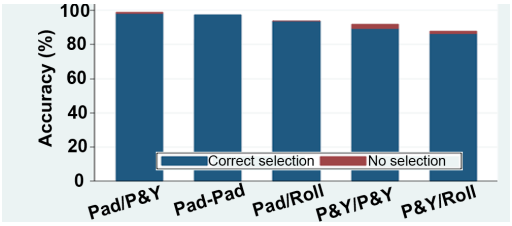

Fig. 7: Mean accuracy rate (% correct) for each technique.

Accuracy (Fig.7). There was a significant effect of technique ($F_{4,48}$= 17.52, p<.0001). Post-hoc Tukey tests showed that Pad/P&Y (98.4%) and Pad/Pad (97.4%) were significantly more accurate than P&Y/P&Y (89.5%) and P&Y/Roll (86.8%). Pad/Roll (94.2%) was also significantly more accurate than P&Y/Roll. Inaccuracy reflected not only erroneous selections but also, in some instances, failure to make any selection. This was especially true of P&Y/P&R (2.5% of total selections; 24.5% of inaccuracies) and P&Y/Roll (1.3% and 10.2%).

Subjective preferences. In the post-study questionnaire the participants ranked the five techniques as follows: Pad/Pad (mean rank=1.5), Pad/P&Y (2.4), P&Y/P&Y (2.8), Pad/Roll (3.5) and P&Y/Roll (4.8). We also recorded subjective evaluations of the 5 techniques using a 7-point Likert scale (Table 4) and analyzed them with a Kruskal-Wallis test. The results showed that Pad/Pad was significantly perceived faster, more accurate, simpler and less tiring than Pad/Roll and P&Y/Roll. They also showed that Pad/P&Y was perceived faster than P&Y/Roll and P&Y/P&Y, and simpler than P&Y/Roll. No technique was judged significantly more 'fun'.

Table 4: Subjective preference

	Pad/Pad	Pad/P&Y	Pad/Roll	P&Y/P&Y	P&Y/Roll
Speed	6.4	5.5	4.5	4.8	3.1
Accuracy	6.8	5.4	4.5	4.9	3.8
Pleasant	6.5	5.5	4.1	5.5	3.3
Tiring	1.2	2.6	4.5	3.2	5.4
Easy To Learn	6.9	5.9	4.8	5.8	3.6
Fun	4.1	5.1	4.3	5.2	4.1

5.4 Discussion

Button-based techniques. The strategy of merging button and gestural interactions was remarkably successful for activating commands. Besides Pad/Pad, Pad/P&Y and, to a lesser extent, Pad/Roll were the fastest techniques (resp. 1.40s, 1.39s and 1.65s) and demonstrated high accuracy rates (resp. 97.4%, 98.4% and 94.2%). This result is worth noting since Pad/Pad was expected to be especially efficient as the user just needs to click two times on the appropriate buttons of the d-pad. We expected mid-air gestures to take more time and be more error prone than purely button-based interactions but we obtained similar results for Pad/Pad and Pad/P&Y, this showing the effectiveness of combining modalities. While slightly inferior Pad/Roll performance was also pretty good.

An important advantage of these three techniques is that they are compatible together: all of them can coexist without interference, thus increasing the number of possible commands. Besides, d-pads are commonly used for other features such as channel or system navigation. This makes the Pad/Pad technique hardy usable: either an additional d-pad would be needed or a button mode that would make the interaction unnecessarily complex. However, this problem does not affect hybrid strategies combining button and gestural interactions. Clicking a button without moving would then produce the usual action when the button is released, while pressing a button then performing a pitch&yaw or a roll gesture would activate a command. Gestural interaction would then provide a mean to augment remote controls while allowing usual interactions.

Purely gestural techniques. P&Y/P&Y gave good results (1.68s, 89.5%) but P&Y/Roll performance (2.02s, 86.8%) was a bit deceiving. A possible explanation is that combining pitch&yaw and roll gestures requires more attention than other gestures. Biomechanically constraints may be another reason, making it hard for users to perform pitch&yaw then roll gestures. For instance, one participant said: "P&Y/Roll would be my favorite technique if there were only three different angular variations" (one of them being the central position, not used in this experiment). Besides, an analysis of the video record revealed that a noticeable number of errors corresponded to correct gestures that were wrongly recognized by the system (about 50% for

P&Y/P&Y and 17% for P&Y/Roll). Better signal processing algorithm, such as those developed by Movea/Gyration [11] would certainly improve results. The problem is that gyroscopes accumulate drift over time and produce errors with motion, even for relatively small movements. As a consequence, the raw data generated by a gyroscope cannot be directly processed by a recognition algorithm. Not only the gyroscope must be recalibrated as often as possible (using an accelerometer) but the data must be constantly corrected (typically by means of Kalman filters). Unfortunately, the algorithm we used, based on trigonometry functions, was not as sophisticated. Hence our accuracy figures should be seen as a lower baseline.

5.4.1 Further Improvements

Besides better gyroscope signal processing, another improvement would be to associate translations and rotations. Study 1 showed that most users naturally perform rotations. Yet, some users preferred translations. Taking both translations and rotations into account would allow more flexibility and would probably improve accuracy. Other specific aspects, such as the fact that users sometimes start the second gesture while the device is already in movement should also be taken into account to finely tune the recognition algorithm and increase its robustness.

We only considered four item menus, mainly because of technical constraints (the Wiimote d-pad provides only four directions) but also to avoid making the experiment too long. Difficulties were reported in GesText [13] for performing diagonal gestures, but contrary to the recommendations that emerged from Study 1 this technique was using pitch&roll rather than pitch&yaw gestures. More work is anyway needed to evaluate the efficiency of pitch&yaw gestures with eight items.

The expert mode involves gesture recall from user memory. Full expert mode evaluation thus requires a memorization and recall phase [20][1]. This means long-lasting experiments out of reach in the present study with many techniques to evaluate. We thus used a simpler experiment where the stimulus indicates which gestures must be performed. However, given the proximity between Marking menus and the techniques based on pitch&yaw or pad gestures, there are likely to share similar properties with respect to memory coding. The case is different with Roll gestures, which requires further research (but this is debatable as they also rely on a circular representations).

6. NOVICE MODE

The proposed techniques are inspired by Marking menus [1][18] [39] and thus provide both an expert and novice mode. The expert mode consists of one of the five above-described techniques and can be performed without visual feedback. The novice mode consists in navigating in a two-level menu which graphical design. It depends on the technique used in expert mode, as explained below. The items of the first level are either activated by Pad or P&Y gestures (depending of the expert mode) and the items of the second level by Pad, P&Y or Roll gestures. Interestingly, hybrid techniques combining button interaction and mid-air gestures allow parallelism as item selection and rely on different modalities for each level.

The novice mode is activated by pressing a button without performing gestures during 300ms. The first-level menu, which is always circular, then appears (Fig. 8, left). Each item corresponds to a submenu, which is also circular except for Roll gestures, in which case it is half-circular (Fig. 8, right). Circular menus can contain up to 8 items. Half of them (on-axis items) will be used in case of a 4 directional d-pad or if diagonal P&Y gestures are disabled. Items are selected in the same way in expert and novice modes. Half-circular menus work according to the same principle but only contains 5 items.

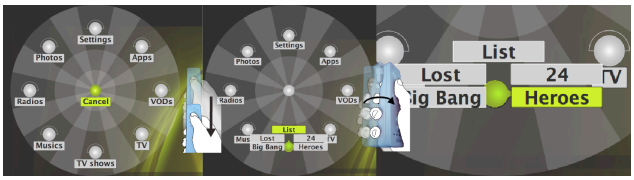

Fig. 8: Novice-mode for P&Y and Pad gestures (left) and Roll gestures (right).

As for regular Marking menus, users can easily understand how the technique works as there is a direct mapping between the orientation of the displayed items and the gestures users have to perform with the thumb or the wrist. A key point of the technique is that it favors a fluid transition from novice to expert usage [18]: as users perform the same gesture in both modes, the user can implicitly learn the expert mode just by repeating the same gesture in novice mode. Finally, our technique not only makes it possible gestural interaction with a remote control, but makes actions visible, easily discoverable and easy to learn, thus answering the concerns expressed in [25] about purely gestural interfaces.

Fig. 9: Yaw gestures for navigating in a table.

Finally, as circular menus can only contain a limited number of items, we also considered the case when activating an item of the first-level menu does not open a submenu but a tabular representation containing as many elements as needed. For instance, selecting the "TV Shows" item in the first-level menu would display the table illustrated in Fig. 9. Roll gestures or d-pad navigation are then used for selecting an item in the table that is currently displayed. Yaw gestures allow making the previous or following table to appear.

7. CONCLUSION

In this article, we investigated the new possibilities offered by mid-air gestures for augmenting and improving remote control interaction. Remote controls tend to be cluttered with buttons, a consequence of an overwhelming number of functions. Mid-air gestures seem appropriate for alleviating this problem while offering the ability to perform favorite actions eyes-free.

The first study revealed that the optimal mapping between mid-air gestures and directional actions strongly depends on context and that pitch & yaw rotations should be preferred for remote control interaction. The second study, focused on the ability of users to perform roll gestures eyes-free, suggested that users can accurately select up to 5 items, up to 7 items with an adaptive algorithm. We then proposed and experimentally compared five different techniques using either mid-air gestures, directional-pad manipulation, or a combination of both.

The results showed that mid-air gestures have promise as an additional input resource. Hybrid techniques combining mid-air gestures and buttons were especially efficient, with the further

advantage of compatibility with pure button-based techniques. Techniques using the d-pad or pitch and yaw for transposing 2D Marking menus to the mid-air space proved especially well suited for eyes-free interaction. Finally, purely gestural techniques yielded slightly lower performance but more careful analysis pointed to a possible bias due to technical limitations.

In this article, we focused on an interesting and little studied context: remote control interaction for controlling interactive television or media-center. We plan to generalize these results to different technologies such as free-hand interaction or different contexts such as large-screen displays.

8. ACKNOWLEDGMENTS

This work was partly achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation as well as the Alexander von Humboldt Fellowship.

9. REFERENCES

[1] Bailly, G., Lecolinet, E., Nigay, L. 2008. Flower menus: a new type of marking menu with large menu breadth, within groups and efficient expert mode memorization. ACM AVI'08, p. 15-22.

[2] Balakrishnan, R., Baudel, T., Kurtenbach, G., Fitzmau-rice, G. 1997. The Rockin'Mouse: integral 3D manipu-lation on a plane. ACM CHI '97, p. 311-318.

[3] Barkhuus, L. and Brown, B. 2009. Unpacking the television: User practices around a changing technology. ACM TOCHI. 16, 3, p. 1-22.

[4] Baudel, T. and Beaudouin-Lafon, M. 1993. Charade: remote control of objects using free-hand gestures. Commun. ACM 36, 7 (Jul. 1993), 28-35.

[5] Cao, X. and Balakrishnan, R. 2003. VisionWand: interaction techniques for large displays using a passive wand tracked in 3D. ACM UIST '03, p.173-182.

[6] Cesar, P., and Chorianopoulos, K. 2009. The evolution of TV systems, content and users toward interactivity. New Foundation and Trends in HCI, vol.2: n°4, p.373-95.

[7] Cho S-J., Oh J. K., Bang W.-C., Chang W., Choi E., Jing Y., Cho J., Kim D. Y. 2004. Magic wand: a hand-drawn gesture input device in 3-D space with inertial sensors. *Frontiers in Handwriting Recognition. IWFHR-9 2004.* p. 106- 111.

[8] Cooper W. 2008. The interactive television user experience so far. Proc. ACM UXTV '08, p. 133-142.

[9] http://sourceforge.net/projects/darwiin-remote/

[10] Freeman, W. T., and Weissman, C. 1995. Television control by hand gestures. Int. Workshop IEEE FG'95, p. 179-183.

[11] Gyration. http://www.gyration.com/

[12] Hogan, N., 1985, The mechanics of multi-joint posture and movement control, Biol. Cybern., 52:315-331.

[13] Jones, E., Alexander, J., Andreou, A., Irani, P., Subra-manian, S. 2010. GesText: accelerometer-based gestural text-entry systems. ACM CHI'10, p. 2173-82.

[14] Kela, J., Korpipää, P., Mäntyjärvi, J., Kallio, S., Savino, G., Jozzo, L., and Marca, D. 2006. Accelerometer-based gesture control for a design environment. Personal Ubiquitous Comput. 10, 5 (Jul. 2006), 285-299.

[15] Kiger. J. I. 1984. The depth/breadth trade-off in the design of menu-driven user interfaces. Int. J. Man-Mach. Stud. 20, 2 (March 1984), 201-213.

[16] Kinect: http://www.xbox.com/kinect

[17] http://en.wikipedia.org/wiki/K_nearest_neighbor_algorithm

[18] Kurtenbach, G. and Buxton, W. 1991. Issues in combining marking and direct manipulation techniques. ACM UIST '91, 137-144.

[19] Kurtenbach, G. and Buxton, W. 1993. The limits of expert performance using hierarchic marking menus. IFIP INTERACT '93 and ACM CHI '93, p. 482-487.

[20] Kurtenbach, G. P., Sellen, A. J., and Buxton, W. A. 1993. An empirical evaluation of some articulatory and cognitive aspect of marking menus. Hum.-Comp. Int.. 8, 1 (Mar. 93), 1-23.

[21] http://www.l5remote.com/

[22] http://www.logitech.com/en-us/remotes/universal-remotes

[23] http://en.wikipedia.org/wiki/Cross-validation_(statistics)

[24] Nelson, M. 1999. Remote controls. The Digital Knowledge Handbook, 3(4).

[25] Norman D. A. 2010 Natural User Interfaces are Not Natural. ACM magazine Interactions 17,3, p. 6-10.

[26] Oakley, I. and Park, J. 2007. Designing eyes-free interaction. Springer HAID'07, p. 121-132.

[27] Oakley, I. and Park, J. 2009. Motion marking menus: An eyes-free approach to motion input for handheld devices. *Int. J. Hum.-Comput. Stud.* 67, 6, p. 515-532.

[28] Partridge, K., Chatterjee, Sazawal, V., Borrielo, G. 2002. TiltType: Accelerometer-Supported Text Entry for Very Small Devices. ACM UIST'02, p. 201-204.

[29] Rahman, M., Gustafson, S., Irani, P., and Subramanian, S. 2009. Tilt techniques: investigating the dexterity of wrist-based input. ACM CHI '09, p. 1943-1952.

[30] Rico, J. and Brewster, S. 2010. Usable gestures for mobile interfaces: evaluating social acceptability. ACM CHI, p. 887-896.

[31] Sazawal, V., Want, R., and Borriello, G. 2002. The Unigesture Approach. Springer Mobile HCI '02, p. 256-270.

[32] Shoemaker, G., Findlater, L., Dawson, J. Q., and Booth, K. S. 2009. Mid-air text input techniques for very large wall displays. ACM GI'09, p. 231-238.

[33] http://www.uwand.com/

[34] Wang, J., Zhai, S., and Canny, J. 2010. SHRIMP: solving collision and out of vocabulary problems in mobile predictive input with motion gesture. ACM CHI '10, p. 15-24.

[35] Wigdor, D. and Balakrishnan, R. 2003. TiltText: using tilt for text input to mobile phones. ACM UIST '03, p. 81-90.

[36] Wilson, A. and Shafer, S. 2003. XWand: UI for intelligent spaces. ACM CHI '03, p. 545-552.

[37] Wobbrock J. O., Morris M. R., Wilson A. D. 2009. User-defined gestures for surface computing. ACM CHI'09, p. 1083-1092.

[38] http://www.voomote.tv/

[39] Zhao, S. and Balakrishnan, R. 2004. Simple vs. compound mark hierarchical marking menus. ACM UIST '04, p. 33-42