How Can I Interact? Comparing Full Body Gesture Visualizations

Felix Kistler
Human Centered Multimedia,
Augsburg University
Universitätsstr. 6a
86159 Augsburg, Germany
kistler@hcm-lab.de

Elisabeth André
Human Centered Multimedia,
Augsburg University
Universitätsstr. 6a
86159 Augsburg, Germany
andre@hcm-lab.de

CHI-PLAY 2015, October 03–07, 2015, London, United Kingdom.
ACM 978-1-4503-3466-2
http://dx.doi.org/10.1145/2793107.2810299

Abstract
This paper is dedicated to the question "How can I interact?", which may arise during a full body interaction game. To answer this question, a game needs to tell the players what actions are available and how those actions can be triggered. We focus on the video channel and use onscreen symbols to visualize how available input gestures have to be performed. We describe three symbol variants using recordings of a real person: color images, tracking shapes and skeletons, and solely tracking skeletons. An initial evaluation study shows clear advantages for the color images. We further outline how we extend the current implementation, for both improving the usability of the symbols, as well as easing their development.

Author Keywords
Full Body Interaction; Gesture; Visualization

ACM Classification Keywords
H.5.2 [Information interfaces and presentation (e.g., HCI)]: User Interfaces.

Introduction
After the release of the Microsoft Kinect[1], full body interaction has become more and more popular. Nevertheless, even a well-designed and robustly

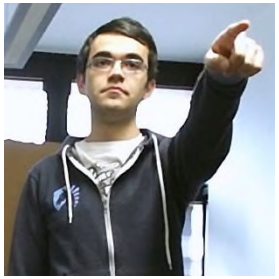
[1] http://www.xbox.com/kinect

implemented full body interaction system can still be difficult to interact with for the actual user. Full body interaction is still quite novel to most users and it inherits many differences to traditional input modalities regarding affordances and responses. For example, users cannot start the interaction as explicitly as with picking up a controller or mouse, and triggering actions is not as simple as pressing a physical button that is part of the interaction device. Therefore, it is important to provide mechanisms for supporting users with the interaction. Along the question "How can I interact?", our goal is to tell the users what actions are available at a specific point in time and how they can be triggered. The most common output channel for such information is video. Therefore, we will look at different types of full body gesture visualizations included in the game interface as onscreen symbols, as in the intercultural learning game Traveller [3] in Figure 1.

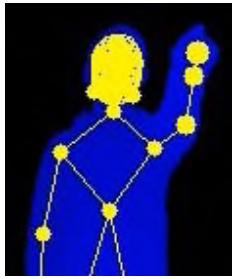
Figure 1: Cyan gesture symbols in Traveller [3]

Such visualizations display a humanoid body or body part performing the gesture, and can differ between displaying the color image of a real person (cf. [8, 1, 6]), abstracting it by its shape or skeleton (cf. [7, 3]), illustrating it with a cartoon-like image such as a virtual character or a schematic drawing (cf. [4] and most commercial Kinect games), or employing a non-human body such as a robot, an animal, or other humanized objects. Color images have the advantage that they are closest to real-life and should be understood easily, while more abstract visualizations simplify in different degrees to let the user concentrate on important information. The abstract images further leave more freedom in design and, especially for commercial games, have aesthetic reasons. Static postures can be visualized by single images of the body in pose, optionally highlighting important parts. For dynamic gestures, there are multiple options to visualize motions. Most common is to animate the image, i.e. playing it as a video. Another option is to add lines or arrows for emphasizing motion trajectories (cf. [1, 3, 5] and most commercial Kinect games). More artistic options, e.g using motion blur, quiver lines, or double takes (cf. [2]), are very rare in literature as well as in commercial games.

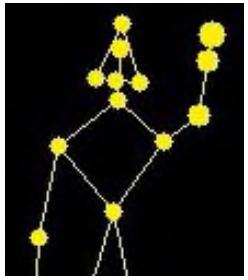
A special focus on teaching full body gestures thoroughly, while correcting users when necessary, can be found in the area of physical training and rehabilitation. For example, Anderson et al. [1] play a color video continuously or pose by pose, while wrong positioned joints are marked with red circles on the users' tracked skeleton. Ribbons leading out of the skeleton joints visualize upcoming movements. The users get an accuracy score and can compare their recorded performances with the training video. Velloso et al. [6] also display color images and the tracked skeleton of the user. In addition, indicators that look like traffic lights visualize whether arm movements are going in the right or wrong direction, a label flashes up in case the movement is performed too fast or slowly, and another label displays the current number of repetitions. Tang et al. [5] display the live color image of the user and try to guide towards the gesture. They render a 2D or 3D arrow at the hand pointing in the desired direction and add a line to show the upcoming path. Alternatively, 2D or 3D lines visualize a trace-ahead of the upcoming whole arm



(a) Color



(b) Shape



(c) Skeleton

Figure 2: Pointing gesture visualized with three different techniques

movements. Zhao et al. [9] display a virtual character exemplifying the exercise, while the user controls a similarly looking avatar next to it, which is similar to Kinect fitness games such as "Your Shape: Fitness Evolved". Zhao et al. further display spheres as joint targets that change their color when they are reached and are tagged with the number of repetitions.

Overall, only few approaches actually compare multiple visualization techniques of full body interaction. Tang et al. [5] perform an informal study with their (feedforward) movement visualizations, which indicates that the 3D versions of arrows and lines make it easier to perceive directions than the 2D versions, while both types still need improvements regarding depth perception. Anderson et al. [1] show that their training system results in better short-term retention scores than traditional video-based instructions. Walter et al. [7] investigate a public display setting. They conclude that more users execute the gesture of a symbol in a dedicated area, instead of one surrounded by other content. Interrupting the current display and showing the symbol alone covering the full screen, makes more users stop the interaction and leave instead. In general, designers of full body interaction apply a certain visualization technique according to their preferences (e.g. color videos in [8] or virtual mannequins in [4]), but they do not investigate further options, and a comprehensive formal evaluation is still missing.

Initial Study

In an initial study we focused on recordings of real persons performing the desired gesture. This is a common practice in scientific studies and eases the creation of gesture symbols. Nevertheless, there are many options on how to visualize the gestures when using recordings of a depth sensor. We investigated three techniques. In the"Color" technique (cf. Figure 2a), a gesture performance was captured on color image as done by Zafrulla et al. [8]. In the "Shape" technique (cf. Figure 2b), the actor was abstracted by only displaying a uni-colored shape enhanced with the Kinect's tracking skeleton and face mesh equal to Kistler et al. [3]. This was further simplified in the "Skeleton" technique (cf. Figure 2c), in which only the Kinect's tracking skeleton with a simplified face was used. The latter might be closest to the schematic drawings used in some commercial Kinect games, e.g. "Kinect Adventures" or "Dragonball Z", but the symbols did not involve manual design. For all three techniques, dynamic gestures were animated as a video with 25 frames per second, but no other enhancements, such as lines or arrows, were introduced. All visualizations had in common that they were generated automatically out of a user's recorded gesture performance, and they only differed in which part of the sensor information they presented.

Regarding the three visualizations, the hypothesis was that "Color" should make it the most easy for users to reproduce the gestures in an accurate way. The reason is that this technique represented the gestures as they are seen in real-life on other people, although without stereoscopic vision. The other two techniques simplified the visualization and only displayed the information used by the tracking system. As those techniques therefore omitted presumably unnecessary details, we assumed that users were faster in starting to perform the gestures. Nevertheless, as the "Skeleton" technique completely abstracted from the actual human shape, it might be again more difficult to translate back to the actual body movement in comparison to the "Shape" technique.

Eighteen gestures were chosen for the study, while trying to cover different body parts, dynamic and static gestures,

and different complexities, e.g. “shaking the head”, “crossing arms”, “walking in place”, or “holding the right leg as if showing an injury of the knee”. “Drawing a circle in the air” further served as a tutorial gesture.

Setup, Procedure and Participants
The experiment was arranged in a room of about 3 meters width and 6.5 meters depth. The participants were standing at a distance of about 2 meters in front of a 50 inch plasma display, with a Kinect for Windows 2 placed just below the screen in a horizontally centered position. The experimenter was sitting to the left of the participant and controlled the application via mouse and keyboard.

After a demographic questionnaire the experimenter explained the study procedure. At first, a timer on the screen was counting down from five to zero. At zero, the first gesture symbol was displayed on the screen. The participants should look at the symbol, and as soon as they understood how the gesture needed to be performed, they should immediately start performing the gesture. However, the gesture performance itself should be done as precise as possible, without trying to be fast. Moreover, the participants were told that the animated gesture symbols were repeated infinitely, but they should start the gesture as soon as one iteration had been played and they understood how to perform it. As soon as the system recognized the gesture performance, the symbol was blended out and the count down timer started again for the next gesture. All participants saw all three types of visualizations, however, their order was counterbalanced. For each visualization, the participants first saw the tutorial gesture and then six gestures of the gesture set, i.e. gestures 2–7 in the first visualization, gestures 8–13 in the second visualization, and gestures 14–19 in the third visualization. Therefore, different participants saw a different subset of the 18 gestures for one visualization type. The symbols were alternating single images and animated videos corresponding to static and dynamic gestures. After each visualization, the participants filled in a short questionnaire regarding usability and intuitiveness of the visualization technique. Further, the participants should name gestures that had been especially hard or easy to reproduce.

During the study, the program recorded the Kinect’s depth, color, and tracking streams. It further measured the time from the symbol display to the successful gesture performance, either using the results of the FUBI real-time recognition system [3] or alternatively, to the experimenter pressing space bar. These timings were synchronized to the recorded video streams and later enhanced manually with the time from the symbol display to the start of the gesture’s preparation phase, i.e. the first movement introducing the gesture performance.

Eighteen participants including two females took part in the study. Their age ranged from 22 to 33 with an average of 26.39 (SD 2.79). All except for two were right-handed. The participants were recruited from our university campus and all had a computer science background. They also stated themselves a medium experience with body gesture based interaction of 2.94 (SD 1.06) on a scale from 1 (no experience) to 5 (practically daily usage).

Results
Regarding the usability and intuitiveness questionnaires, we calculated an overall score in the range of 1–5 for each visualization and participant as the average of all responses. A one-way repeated measures ANOVA indicated a significant effect with $F(2, 17) = 9.59$, $p < 0.001$, $\eta^2 = 0.36$. Post-hoc tests with Bonferroni correction showed that the “Color” technique was rated

significantly better than the other two techniques. "Color" reached a mean overall score of 4.78 (SD 0.40) being significantly higher ($p < 0.01$) than "Shape" and "Skeleton" ($r_{Color\_Shape} = 0.65$, $r_{Color\_Skeleton} = 0.68$). However, there was no significant difference ($p > 0.05$) between the mean score of "Shape" with 3.97 (SD 0.99) and the one of "Skeleton" that reached 3.56 (SD 1.20).

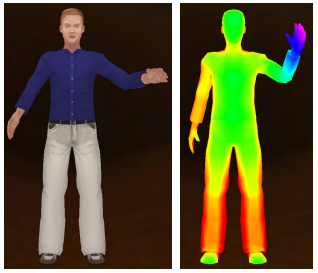
We further looked at the time it took the participants to start the gesture performance after the symbol had been displayed, using the annotated timings.
Greenhouse-Geisser corrected estimates were significant with $F(1.27, 21.61) = 5.01$, $p < 0.05$. Participants were faster (AVG 1.60 sec, SD 0.38)in starting the gesture with the "Color" technique than with the other two techniques ($p_{Color\_Shape} < 0.01$, $r_{Color\_Shape} = 0.68$, $p_{Color\_Skeleton} < 0.05$, $r_{Color\_Skeleton} = 0.57$).
Nevertheless, there was again no significant difference ($p > 0.05$) between "Shape" that reached 1.86 sec (SD 0.42) and "Skeleton" with 1.97 sec (SD 0.72).

Finally, we compared the number of wrongly performed gestures (="wrong") as well as the number of gestures rated as especially hard to reproduce (="hard"). For both, Friedman ANOVAs reported a significant effect, $p < 0.001$, $\chi^2_{wrong}(2) = 16.45$, $\chi^2_{hard}(2) = 12.81$, but only the comparison between "Color" and "Skeleton" was significant ($p < 0.0167$ for Bonferroni correction, $r_{wrong} = -0.58$, $r_{hard} = -0.55$), while the comparisons between "Shape" and the other two techniques were not. "Color" had no wrongly performed gestures and an average of 0.22 (SD 0.43) gestures were rated as hard to reproduce. "Shape" had an average of 0.33 (SD 0.49) wrongly performed gestures and 0.78 (SD 0.73) gestures were rated as hard to reproduce. "Skeleton" had an average of 0.78 (SD 0.55) wrongly performed gestures and 1.00 (SD 0.59) gestures were rated as hard to reproduce.

Conclusion and Discussion
We presented three techniques for visualizing full body gestures, generated automatically by recording the gesture performances with a depth sensor. The techniques were tested in a study with 18 sample gestures preformed by 18 participants. During the study, videos of the participants were recorded and automatically pre-annotated using the FUBI real-time gesture recognition system [3]. We found clear preferences for using color images of real persons to visualize full body gestures. This can be seen in the subjective participants' ratings as well as in the objectively measured time it took them to start performing the gesture and the number of wrongly performed gestures.

The reason that it was easier for the participants to learn the correct performance might be that the color images represented the gestures in a similar way they would be perceived in real life. According to the comments of the participants, the other more abstract gesture visualizations made it especially hard to recognize how the joints should be positioned in depth.

While many scientific approaches already use color images for gesture visualizations, the others should reconsider whether they have good reasons to use abstractions. Using recordings of a real person is common practice in research, as it allows to rapidly create gesture symbols without advanced design skills. However, this practice is almost never applied in commercial games, probably for aesthetic reasons. As we did not investigate the more synthetic visualizations that are commonly used in the gaming industry, the findings of our study do currently not allow clear recommendations for commercial games.



(a) Virtual Character (*left:* textured; *right:* depth colorized)

(b) Active Limb Highlighting

(c) Corrective Arrows

Figure 3: Gesture Visualizations with a Virtual Character

Enhancements and Future Work

To better investigate the visualizations commonly used in commercial games, we implemented gesture symbols based on virtual characters in a next step (cf. Figure 3). Using a (textured) virtual character was closest to the favored "Color" technique, but still had the option to be changed or enhanced later. As a major problem of the symbols was missing depth perception, we added a new shading technique similar to the "Shape" technique, but encoding depth information with colors, from magenta= near to red=far (cf. right-hand image of Figure 3a). Finally, we enhanced the gesture symbols with information similar to the approaches for physical rehabilitation and training [1, 6, 5, 9]. For example, we added highlighting of limbs involved in a gesture (cf. Figure 3b) or included arrows depicting how to adapt joint movements for correctly performing a gesture (cf. Figure 3c).

In the study presented in this paper, we utilized the FUBI real-time recognition system [3] to pre-annotate the gesture performances, using FUBI's XML-based gesture definition language. Currently, we are working on automatically generating visualizations out of the gestures' XML definitions with a virtual character and a limited set of parameters. This should make the creation of the gesture symbols easier, while it will still leave the flexibility to change the design at a later stage. Another option to improve the depth perception could be the use of a stereoscopic display, although this would be harder to setup in real-life, and it will probably be an option only for applications that are already using stereoscopy. Other visualization options, e.g. schematic drawings, could make the perception easier while still being relatively abstract, but would require more manual adaptations. After finding favorite visualization types, they should as well be tested in a real game context, to ensure their practical benefit.

References

[1] Anderson, F., Grossman, T., Matejka, J., and Fitzmaurice, G. YouMove: Enhancing movement training with an augmented reality mirror. In *Proc. UIST 2013*, ACM (New York, 2013), 311–320.

[2] Guigar, B. *The Everything Cartooning Book: Create Unique And Inspired Cartoons For Fun And Profit*. Everything Books, 2004.

[3] Kistler, F., and André, E. User-defined body gestures for an interactive storytelling scenario. In *Human-Computer Interaction – INTERACT 2013*, Springer Berlin Heidelberg (2013), 264–281.

[4] Portillo-Rodriguez, O., Sandoval-Gonzalez, O., Ruffaldi, E., Leonardi, R., Avizzano, C., and Bergamasco, M. Real-time gesture recognition, evaluation and feed-forward correction of a multimodal Tai-Chi platform. In *Haptic and Audio Interaction Design*, Springer Berlin Heidelberg (2008), 30–39.

[5] Tang, R., Alizadeh, H., Tang, A., Bateman, S., and Jorge, J. A. Physio@Home: Design explorations to support movement guidance. In *Proc. CHI EA 2014*, ACM (New York, 2014), 1651–1656.

[6] Velloso, E., Bulling, A., and Gellersen, H. MotionMA: Motion modelling and analysis by demonstration. In *Proc. CHI 2013*, ACM (New York, 2013), 1309–1318.

[7] Walter, R., Bailly, G., and Müller, J. StrikeAPose: Revealing mid-air gestures on public displays. In *Proc. CHI 2013*, ACM (New York, 2013), 841–850.

[8] Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., and Presti, P. American sign language recognition with the Kinect. In *Proc. ICMI 2011*, ACM (New York, 2011), 279–286.

[9] Zhao, W., Feng, H., Lun, R., Espy, D., and Reinthal, M. A Kinect-based rehabilitation exercise monitoring and guidance system. In *Proc. ICSESS 2014* (2014), 762–765.