Traveller: Eine interaktive virtuelle Umgebung für kulturelles Lernen, steuerbar mit benutzerdefinierten Ganzkörpergesten

Felix Kistler & Elisabeth André
Lehrstuhl Human Centered Multimedia, Universität Augsburg,
Universitätsstr. 6a, 86159 Augsburg, Deutschland
kistler@informatik.uni-augsburg.de

Zusammenfassung. In diesem Beitrag beschreiben wir eine virtuelle Umgebung für kulturelles Lernen, welche auf einer interaktiven Geschichte aufbaut. Um die darin enthaltene Interaktion zu verbessern, führten wir eine Studie zur Erzeugung eines benutzerdefinierten Gestensets durch. 22 TeilnehmerInnen durchliefen die Anwendung, wobei die eigentliche Interaktion deaktiviert war. Stattdessen wurde nur angezeigt, was für Aktionen zum jeweiligen Zeitpunkt verfügbar waren. Die Aufgabe der TeilnehmerInnen war es dann, sich Ganzkörpergesten auszudenken und auszuführen, die sich ihrer Meinung nach gut dafür eignen würden, die jeweilige Aktion auszulösen. Die Gestenausführungen wurden auf Video aufgenommen und später analysiert, um für alle Aktionen Gestenkandidaten zu finden, welche von einer Mehrheit der Nutzer gewählt wurden. Diese Gestenkandidaten sind letztendlich mit Hilfe unseres Frameworks zur Erkennung von Ganzkörpergesten wieder in unsere Anwendung integriert worden.

1. Einleitung

Die Einführung der Microsoft Kinect machte Tiefensensoren für NormalverbraucherInnen verfügbar und bot gleichzeitig auch für virtuelle Lernumgebungen die Möglichkeit, neuartige Ganzkörperinteraktion zu integrieren. Auch wenn einige ForscherInnen dies schon für ihre Anwendungen, basierend auf interaktiven Geschichten, durchgeführt haben (z.B. von Álvarez & Peinado (2012), Kistler et al. (2011), so wird die Wahl der Gesten gewöhnlich durch die EntwicklerInnen selbst getroffen. Allerdings müssen durch die EntwicklerInnen bestimmte Gesten nicht zwangsläufig intuitiv für die Mehrheit der eigentlichen NutzerInnen sein. Einen anderen Ansatz zur Gestenfindung verfolgte Wobbrock et al. (2009) mit der Erstellung von benutzerdefinierten Touchscreen-Gesten, was auch schon in einigen anderen Bereichen adaptiert wurde, z. B. von Kurdyukova et al. (2012) für Public Display Umgebungen oder von Obaid et al. (2012) für die Interaktion mit humanoiden Robotern. Die grundlegende Idee hinter seinem Ansatz ist, dass bestimmte Effekte innerhalb eines Systems BenutzerInnen gezeigt werden, woraufhin es deren Aufgabe ist, sich Gesten zu überlegen und auszuführen, welche ihrer Meinung nach gut dafür geeignet sind, diese Effekte auszulösen. Die Gestenausführungen der verschiedenen NutzerInnen werden aufgezeichnet und später analysiert, um Gestenkandidaten zu finden, die von einer Mehrheit der NutzerInnen gewählt wurden.

In diesem Beitrag beschreiben wir Traveller („Train for Virtually Every Locality“, übersetzt: „Trainieren für nahezu jeden Ort“), eine Anwendung für kulturelles Lernen von jungen Erwachsenen (18 bis 25 Jahre), für welche wir auf ähnliche Weise ein benutzerdefinertes Gestenset erstellt haben. Die NutzerInnen sollen aktiv in der Erzählung unserer Anwendung teilnehmen und interagieren mit synthetischen Kulturen, definiert nach Hofstede (2005). Zur Interaktion stehen Aktionen für Navigation in der virtuellen Welt (*hier*: Änderung von Position und Orientierung) und Dialoge mit virtuellen

Charakteren bereit, welche ausschließlich durch Ganzkörpergesten gesteuert werden sollen.

2. Szenario und Interaktionen

Innerhalb der Geschichte befinden sich die Benutzer auf einer Reise durch verschiedene Länder, in welchen sie je nach der dort simulierten Kultur unterschiedliche Aktionen auswählen müssen, um erfolgreich zu sein. Unsere Gestenstudie wurde für einen ersten Teil der Geschichte durchgeführt, welche zunächst im Café der Großmutter (der Erzählfigur) beginnt und danach an zwei verschiedene Orte im ersten Land der Reise führt. Die Großmutter liefert zunächst eine Einführung in das Szenario. Im ersten Land müssen die BenutzerInnen dann als erstes eine Übernachtungsgelegenheit finden, indem sie mit mehreren Personen in einer Bar interagieren. Danach sollen sie die Erlaubnis zum Eintritt in einen Park erlangen, wofür sie aber zunächst den zuständigen Parkaufseher in einem nahe gelegenen Museum ausfindig machen müssen. Mittlerweile besitzt unsere Anwendung schon eine längere Handlung, unsere Gestenstudie wurde allerdings für diesen ersten Teil durchgeführt, welcher zu der Zeit die folgenden Aktionen enthielt, die letztlich durch Ganzkörpergesten ausgelöst werden sollen (vom Englischen übersetzt): *Bejahen, Verneinen, an die Bar sitzen und abwarten, nach dem Weg fragen, auf die Gruppe zugehen, die Bar verlassen, nach dem Parkaufseher fragen, den Assistenten fragen ob er mit dem Parkaufseher reden könnte, auf den Parkaufseher zugehen,* und *den Parkaufseher um Erlaubnis fragen.* Die Anwendung wurde mit Unity3D[1] und einer Agentenarchitektur für kulturell anpassbare Verhaltensweisen von Dias et al. (2011) implementiert.

3. Gestenstudie

3.1 Aufbau, Ablauf und TeilnehmerInnen der Studie

Unsere Studie fand in einem Raum mit ca. 3 mal 6,5 Meter Bodenfläche statt. Die TeilnehmerInnen standen in einem Abstand von ca. 2,5 Metern vor einem 50 Zoll Plasma Bildschirm. Links vom Bildschirm war eine Kamera auf ungefähr 1,5 Metern Höhe positioniert, um die BenutzerInnen von leicht schräg vorne aufzunehmen. Die BenutzerInnen wurden angewiesen, sich auf die Anfangsposition zu stellen. Es sei ihnen aber trotzdem erlaubt, sich frei im Sichtfeld der Kamera zu bewegen. Der Versuchsleiter saß links der StudienteilnehmerInnen und steuerte die Anwendung mit Maus und Tastatur.

Die BenutzerInnen durchliefen die Anwendung, wobei die eigentliche Interaktion deaktiviert war, auf dem Bildschirm aber immer die aktuell möglichen Aktionen mit ihrem Namen in Textfeldern angezeigt wurden. Für diese Aktionen sollten sich die BenutzerInnen daraufhin Gesten überlegen und ausführen. Den TeilnehmerInnen wurde gesagt, sie könnten die Gesten mit ihrem kompletten Körper ausführen, es sei aber am wichtigsten, dass die Geste für sie möglichst gut zur verknüpften Aktion passe. Außerdem sollte die Geste die Aktion auf irgendeine Weise semantisch darstellen und nicht nur aus Zeigen auf das am Bildschirm angezeigte Textfeld bestehen. Um die Studie reproduzierbar zu halten, nannte der Versuchsleiter immer die nächste Aktion, die die NutzerInnen umsetzen sollten. Nachdem Ausführen der Geste sollten die NutzerInnen auf einer Skala

1 http://www.unity3d.com/

von 0=*sehr schwer* bis 7=*sehr leicht* angeben, wie leicht es für sie war, sich eine Geste zur jeweiligen Aktion zu überlegen.

An der Studie nahmen 22 Personen teil, darunter vier Frauen. Ihr Alter lag zwischen 22 und 35 Jahren mit einem Durchschnitt von 26,23 (σ 3,80). Alle bis auf einen waren RechtshänderInnen und alle hatten schon mindestens anfängliche Erfahrungen mit Interaktion über Körpergesten.

3.2 Ergebnisse

Die Gestenausführungen in den Videos wurden nach der Studie annotiert und in Gruppen von gleichen Gesten eingeteilt. Als Gestenkandidat wurde dann zunächst die Gestengruppe ausgewählt, in welche die Gestenausführung der meisten TeilnehmerInnen fiel. Als alternativer Gestenkandidat wurde zusätzlich noch die Gruppe mit den zweitmeisten Gesten in Betracht gezogen, aber nur falls diese mindestens halb so viele Gestenausführungen beinhaltete wie der favorisierte Kandidat.

Tab. 1 beinhaltet die Gestenkandidaten für die zehn untersuchten Aktionen. In der dritten Spalte steht zusätzlich der Anteil, wie oft dieser Kandidat unter allen Gestenvorschlägen auftrat.

Aktion	Gestenkandidaten	Ausgewählt von
Bejahen	Kopfschütteln	68%
Verneinen	Kopfnicken	68%
An die Bar sitzen und abwarten	Hinsetzen	56%
Auf die Gruppe zugehen	Schritt nach vorne	56%
Nach dem Weg fragen	Arme nach außen	34%
Die Bar verlassen	Wegdrehen	45%
	Schritt nach hinten	27%
Nach dem Parkaufseher fragen	Arme nach außen	50%
Den Assistenten fragen ob er mit dem Parkaufseher reden könnte	Auf einen nach dem anderen zeigen	38%
	Nach vorne zeigen	21%
Auf den Parkaufseher zugehen	Schritt nach vorne	56%
Den Parkaufseher um Erlaubnis fragen	Arme nach außen	23%
	Auf die Schulter tippen	19%

Tabelle 1. Gestenkandidaten der Aktionen

Abb. 1 zeigt die durchschnittliche Bewertung der NutzerInnen, wie leicht es war, sich eine Geste für die jeweilige Aktion auszudenken. Die Fehlerbalken stellen den Standardfehler dar und die Aktionen sind anhand ihrer Bewertung geordnet.

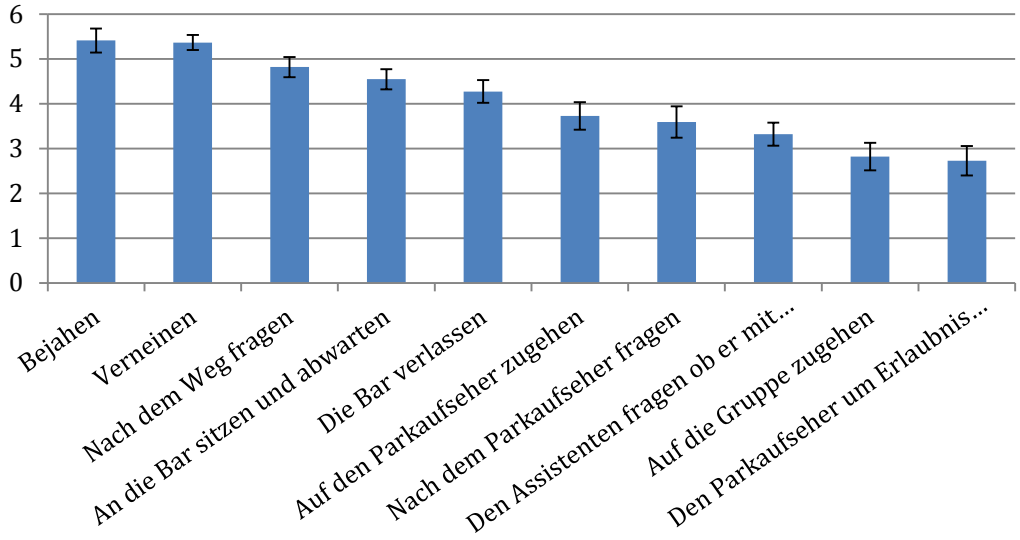


Abbildung 1. Nutzerwertungen für die 10 Aktionen

Eine einfache ANOVA für wiederholte Messungen ergab, dass die Bewertungen sich signifikant zwischen den Aktionen unterschieden $F(9, 21) = 15.90$, $p < 0.01$, $\eta^2 = 0.43$. Post-hoc Tests mit Bonferroni Korrektur zeigten, dass BenutzerInnen es bei Dialog-Aktionen schwieriger fanden, eine Geste zu erfinden. Diese komplexeren Aktionen wurden deshalb signifikant niedriger (p<0,01) bewertet als *Bejahen* und *Verneinen.* Ebenso wurde *auf den Parkaufseher zugehen* und *die Bar verlassen* signifikant (p<0,05) höher gewertet als alle Dialog-Aktionen bis auf *den Parkaufseher um Erlaubnis fragen. Auf die Gruppe zugehen* wurde nur signifkant höher (p <0,05) als *nach dem Parkaufseher fragen* bewertet. Für alle übrigen Paarvergleiche fanden wir keine signifikanten Unterschiede.

Um zusätzlich die Übereinstimmung zwischen den TeilnehmerInnen zu untersuchen, berechneten wir ein Übereinstimmungsmaß (agreement score = AS) basierend auf dem Vorgehen von Wobbrock et al. (2009). Für eine Aktion a, definiert sich die Übereinstimmung $AS(a)$ nach folgender Formel:

$$AS(a) = \sum_{i \in 1..n_a)} \left(\frac{|M_i(a)|}{|M(a)|} \right)^2$$

Die Übereinstimmung $AS(a)$ für eine Aktion a wird folglich durch eine Zahl im Intervall $[1 \,/\, |M(a)|\,, 1]$ repräsentiert, wobei eine größerer Wert auch für eine größere Übereinstimmung steht und 1 eine perfekte Übereinstimmung darstellt (=alle TeilnehmerInnen wählten die gleiche Geste).

Die Gesamtübereinstimmung für die zehn untersuchten Aktionen war 0,329 (σ=0,129). Abb. 2 zeigt die einzelnen Übereinstimmungsmaße für die zehn Aktionen. Diese sind von links nach rechts von größter nach kleinster Übereinstimmung geordnet. Auch hier haben die Dialog-Aktionen die niedrigste Übereinstimmungsbewertung bekommen. Außerdem zeigte sich eine starke Korrelation zwischen dem Übereinstimmungsmaß und den Bewertungen, wie schwierig das Erfinden von Gesten war (Pearson's $r = 0.812, p < 0.01$). Falls BenutzerInnen es einfach fanden, eine Geste für eine Aktion zu erfinden, wählten auch mehr BenutzerInnen die gleichen Gesten aus. Falls dies als schwieriger empfunden wurde, ergab sich auch eine höhere Anzahl an verschiedenen Gesten.

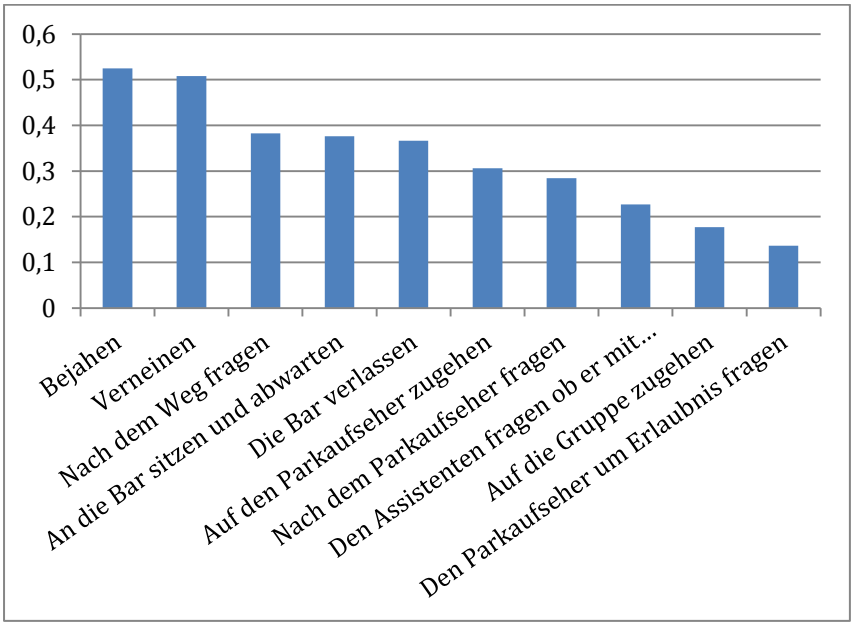


Abbildung 2. Übereinstimmungsmaße für die 10 Aktionen

4. Integration der Gesten

Die gefundenen Gestenkandidaten integrierten wir mit Hilfe unseres Opensource-Frameworks „FUBI“[2] zur Erkennung von Ganzkörpergesten in unsere Anwendung. Eine frühere Version des Frameworks ist in Kistler et al. (2012) beschrieben. In genanntem Framework werden Gesten anhand einer einfachen XML-Syntax definiert.

Die XML Dateien können zunächst drei Typen von Grunderkennern enthalten:

- *Gelenk-Orientierungs-Erkenner* sind durch einen minimalen und/oder maximalen Winkel für ein bestimmtes Gelenk definiert.
- *Gelenk-Relations-Erkenner* betrachten die relative Position eines Gelenkpunktes zu einem anderen, also z. B. ob und wie weit sich eine Hand über der Schulterhöhe befindet oder wie groß der Abstand zum Kopf ist.
- Erkenner für *lineare Bewegungen* sind durch eine bestimmte Bewegungsrichtung und eine minimale und/oder maximale Geschwindigkeit definiert.

Zusätzlich können diese drei Grundtypen in sogenannten *Kombinationserkennern* zu Sequenzen zusammengefügt werden. Ein *Kombinationserkenner* besteht aus mehreren Zuständen, die eine Menge der oben genannten Erkenner referenziert. Für jeden Zustand können eine minimale und maximale Dauer definiert werden, für welche die referenzierten Erkenner erfüllt sein müssen. Zudem kann noch ein Maximum für die Zeit zwischen zwei Zuständen angegeben werden. Abb. 3 zeigt die XML Definition für einen Erkenner von Kopfschütteln.

2 http://www.hcm-lab.de/fubi.html

```
<!--HeadNod-->
<JointOrientationRecognizer name="HeadDown">
  <Joint name="head"/>
  <MaxDegrees x="-13"/>
</JointOrientationRecognizer>
<JointOrientationRecognizer name="HeadUp">
  <Joint name="head"/>
  <MinDegrees x="-5"/>
</JointOrientationRecognizer>
<CombinationRecognizer name="HeadNod">
  <State maxDuration="1" timeForTransition="0.4">
    <Recognizer name="HeadUp"/>
  </State>
  <State maxDuration="1" timeForTransition="0.4">
    <Recognizer name="HeadDown"/>
  </State>
  <State maxDuration="1" timeForTransition="0.4">
    <Recognizer name="HeadUp"/>
  </State>
  <State>
    <Recognizer name="HeadDown"/>
  </State>
</CombinationRecognizer>
```

Abbildung 3. XML Definition zur Erkennung von Kopfschütteln

Die Gesten sind in der Anwendung mit einem Symbol auf dem Bildschirm verknüpft. Ein Symbol kann aus einem einzelnen Bild oder einer Animation bestehen und sobald ein Solches angezeigt wird, überprüft unser Framework, ob die zugehörige Geste von einem Benutzer vor dem Bildschirm ausgeführt wird. Im Erfolgsfall wird daraufhin die zugehörige Aktion ausgelöst. Tab. 1 zeigt die Symbole der Gestenkandidaten, welche wir in unsere Anwendung integriert haben. Die meisten sind tatsächlich Animationen, der Übersichtlichkeit halber wurden aber nur je ein bis zwei Bilder aus der Animation dargestellt.

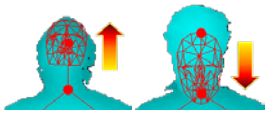
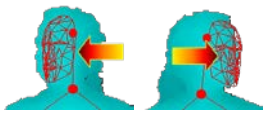
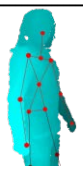
 Kopfnicken *(Bejahen)*		 Kopfschütteln *(Verneinen)*		 Hinsetzen *(an die Bar sitzen und abwarten)*
 Schritt nach vorne *(auf die Gruppe zugehen)*	 Weg drehen *(die Bar verlassen)*	 Nach vorne zeigen *(den Assistenten fragen ob er mit dem Parkaufseher reden könnte)*	 Auf die Schulter tippen *(den Parkaufseher um Erlaubnis fragen)*	 Arme nach außen/Schultezucken *(nach dem Weg fragen, nach dem Parkaufseher fragen)*

Tabelle 2. Symbole der implementierten Gestenkandidaten (zugehörige Aktionen in Klammern)

Abb. 1 zeigt die Barszene im ersten besuchten Land mit vier aktuell zur Auswahl stehenden Aktionen, die durch ihre Symbole des neuen Gestensets dargestellt werden.

Abbildung 4. Gestensymbole am Anfang der Barszene

5. Diskussion

Wir haben vorgeschlagen, die Gestenkandidaten danach auszusuchen, wie viele NutzerInnen eine Geste für eine Aktion gewählt haben. Das muss jedoch nicht zwangsläufig die beste Wahl sein. Es kann vorkommen, dass für mehrere parallele Aktionen die gleiche Geste gewählt wurde. Hier kann es sinnvoll sein, einen weniger oft gewählten Kandidaten zu nehmen, um die Gesten unterscheidbar zu halten. Genauso könnte es sein, dass eine von BenutzerInnen favorisierte Geste technisch nicht umgesetzt werden kann oder dass symmetrische Aktionen (z.B. nach links gehen und nach rechts gehen) besser auch durch symmetrische Gesten repräsentiert werden sollten, um eine konsistentere Interaktion zu garantieren.

Ein grundsätzliches Problem in unserem Szenario waren die komplexeren Dialog-Aktionen, welche niedrigere Übereinstimmung bei der Gestenwahl und auch eine schlechtere Benutzerwertung bekamen. Zum Zeitpunkt der Studie enthielt unser Szenario nie mehrere Dialog-Aktionen gleichzeitig, weshalb deren mehrfach überschneidende Gesten kein Problem für uns darstellten. Sobald dies allerdings der Fall ist, müssen wir die Interaktion hier noch weiter anpassen, entweder durch Aufspalten der Aktionen in mehrere Schritte oder durch Einführung einer anderen Eingabemodalität, z. B. mit Interaktion über ein grafisches Benutzerinterface oder über Spracheingabe.

6. Fazit und Ausblick

In diesem Beitrag haben wir eine interaktive virtuelle Umgebung für kulturelles Lernen beschreiben, für welche wir ein benutzerdefiniertes Gestenset bestimmt haben. Dafür wurden verschiedene Gesten untersucht, die von BenutzerInnen spontan zu genannten Aktionen ausgedacht wurden. Die Gesten wurden mit Hilfe unseres

Erkennungsframeworks FUBI (Kistler et al. 2012) am Schluss dann wieder in die Anwendung integriert.

Eine erste Validierung des FUBI Frameworks bezüglich Erkennungsrate und Nutzerzufriedenheit wurde schon mit einem anderen interaktiven Szenario durchgeführt (Kistler et al. 2012). Wir planen, eine erweiterte Validierung mit dem aktuellen Szenario durchzuführen. Des Weiteren werden wir die Gestenfindung für den Rest des Szenarios fortsetzen, sobald dieses weiter entwickelt ist. Außerdem wurde inzwischen klar, dass unser Szenario auch Stellen enthalten wird, in denen eine größere Anzahl komplexer Konversations-Aktionen parallel auftauchen. Dadurch wird es nicht mehr möglich sein, alle Aktionen durch unterschiedliche Gesten sinnvoll zu repräsentieren. Aus diesem Grund habe wir uns dafür entschieden eine zusätzliche Interaktionsart einzubauen, in der mit Handbewegungen in der Luft Elemente in einer grafischen Benutzeroberfläche selektiert werden können.

Danksagungen. Die hier beschriebene Arbeit wurde durch die Europäische Kommission innerhalb des FP7 finanziert, eCUTE (FP7-ICT-257666).

7. Literaturverzeichnis

Álvarez & Peinado (2012) Álvarez, N. & Peinado, F. (2012), Exploring body language as narrative interface, in Interactive Storytelling, Lecture Notes in Computer Science, Springer Berlin Heidelberg, S. 196–201.

Dias et al. (2011) Dias, J., Mascarenhas, S. & Paiva, A. (2011), Fatima modular: Towards an agent architecture with a generic appraisal framework, in Proc. of the Int. Workshop on Standards for Emotion Modeling.

Hofstede (2005) Hofstede, G. J. (2005), Role playing with synthetic cultures: the evasive rules of the game, Experimental Interactive Learning in Industrial Management: New approaches to Learning, Studying and Teaching, S. 49.

Kistler et al. (2012) Kistler, F., Endrass, B., Damian, I., Dang, C. & André, E. (2012), Natural interaction with culturally adaptive virtual characters, Journal on Multimodal User Interfaces 6, S. 39–47.

Kistler et al. (2011) Kistler, F., Sollfrank, D., Bee, N. & André, E. (2011), Full body gestures enhancing a game book for interactive story telling, in Interactive Storytelling, Vol. 7069 von Lecture Notes in Computer Science, Springer Berlin Heidelberg, S. 207–218.

Kurdyukova et al. (2012) Kurdyukova, E., Redlin, M. & André, E. (2012), Studying user-defined iPad gestures for interaction in multi-display environment, in Proc. IUI 2012, S. 1–6.

Obaid et al. (2012) Obaid, M., Häring, M., Kistler, F., Bühling, R. & André, E. (2012), User-defined body gestures for navigational control of a humanoid robot, in Social Robotics, Vol. 7621 von Lecture Notes in Computer Science, Springer Berlin Heidelberg, S. 367–377.

Wobbrock et al. (2009) Wobbrock, J. O., Morris, M. R. & Wilson, A. D. (2009), User-defined gestures for surface computing, in Proc. CHI 2009, ACM, New York, NY, USA, S. 1083–1092.