Hochschule Reutlingen
Reutlingen University

INF
Studiengang
Human-Centered
Computing

Uwe Kloos, Natividad Martínez, Gabriela Tullius (Hrsg.)

Wissenschaftliche Vertiefungskonferenz

huc - Mensch im Mittelpunkt

Informatik-Konferenz an der Hochschule Reutlingen

18. November 2015

ISBN 978-3-00-051706-8

Impressum

Anschrift:
Hochschule Reutlingen
Reutlingen University
Fakultät Informatik
Human-Centered Computing
Alteburgstraße 150
D-72762 Reutlingen

Telefon: +49 7121 / 271-4002
Telefax: +49 7121 / 271-4042

E-Mail: wvk@reutlingen-university.de
Internet: wvk.reutlingen-university.de

Organisationskomitee:
Prof. Dr. Uwe Kloos, Hochschule Reutlingen
Prof. Dr. Natividad Martínez, Hochschule Reutlingen
Prof. Dr. Gabriela Tullius, Hochschule Reutlingen

Ralf Dauenhauer
Peter Einberger
Raphael Fritsch
Thomas Gulde
Sebastian Hirth
Simone Liegl
Lasse Naumann
Alexander Niedel
Paul Pasler
David Randler
Veronika Rein
Steffen Schellig
Lukas Schmitt
Florian Strieg
Dominik Waas

Hochschule Reutlingen
Reutlingen University

Herstellung und Verlag: Hochschule Reutlingen
ISBN 978-3-00-051706-8

Vorwort

Jahrzehnte lang haben wichtige Programme und Anwendungen die Menschheit durch mangelnde Verständlichkeit gequält. Der Nutzer, man könnte aber auch sagen, der Kunde stand vor manchem Programm verärgert und ratlos, da die Computermaschine scheinbar nur zu ihrem Selbstzweck funktionierte. Viele taten sich schwer zu verstehen wie man das Programm bedienen sollte, was das Programm einem nützen könnte. Mittlerweile hat die Informationswissenschaft aber auch den Menschen, den Benutzer, im positiven Sinne den „Kunden“, im Visier: was nutzt das beste Programm, wenn es keinem dient? Aktuell schlägt die wissenschaftliche Forschung und Entwicklung einen Weg ein, der den Nutzen für den Menschen in den Mittelpunkt stellt - Human-Centered Computing. Der Computer soll dem Menschen dienen, nicht umgekehrt. Dies ist seit 2013 auch das Kernthema im gleichnamigen Studiengang - kurz huc - an der Fakultät Informatik in Reutlingen.

Diese Konferenz wissenschaftliche Vertiefung 2015, die im früheren mki-Masterstudiengang eingeführt wurde, jährt sich zum 11. Mal. Der huc Masterstudiengang ist mittlerweile erwachsen geworden. 16 Studentinnen und Studenten stellen in diesem Heft ihre Beiträge zu dem Thema „huc - Mensch im Mittelpunkt“ zur Diskussion. Dabei werden spannende und wichtige Aspekte behandelt, die allesamt dem Menschen, der Gesellschaft ‚dienen‘ sollen: computerunterstützte Optimierung medizinischer Diagnosen (Lasse Naumann), Cloud - Sicherheit (Dominic Waas) oder Risiken und Probleme von One-Time Passwörtern (Steffen Schellig). Sebastian Hirth stellt die Frage nach gutem Produktdesign, z. B. zur besseren Bedienung medizinischer Geräte in Praxen und Krankenhäusern. Auch Zukunftsthemen, die scheinbar „verrückt“ oder visionär klingen, stehen im aktuellen wissenschaftlichen Diskurs, wenn man z. B. mit dem Finger in der Luft, auf holografische Projektionen tippt, um Geräte zu steuern (Alexander Niedel). Eine weitere aktuelle Herausforderung für huc Studierende: der zunehmende Verkehrsstress; Paul Pasler forscht an Methoden zur Müdigkeitserkennung in Fahrzeugen.

Ich wünsche Ihnen viel Freude bei dieser wissenschaftlichen Lektüre. Möge dem einen Leser oder der anderen Leserin diese Themen noch nachhaltig von Nutzen sein.

23. November 2015

Prof. Boris Terpinc

Inhaltsverzeichnis

Erstellung eines Klassifikationsschemas zur Informationsverknüpfung in Augmented-Reality-Anwendungen *

Ralf Dauenhauer
Reutlingen University
Ralf.Dauenhauer@Student.Reutlingen-University.DE

Abstract

Interface Richtlinien stellen eine wichtige Basis für den erfolgreichen Einsatz von Augmented-Reality-Anwendungen zur Unterstützung bei Wartungs- und Reparaturarbeiten dar. Insbesondere über die Art der Verknüpfung von physischen und virtuellen Objekten existieren bisher nur vereinzelte wissenschaftliche Untersuchungen. In dieser Ausarbeitung wird ein neuartiges Klassifikationsschema vorgestellt, mit dem Varianten zur Informationsverknüpfung kategorisiert werden können. In diesem Zusammenhang werden die einzelnen Bestandteile, physischer Anker, Informationsobjekt sowie Metaobjekte, identifiziert und definiert, welche im Darstellungsraum existieren. Das Klassifikationsschema besteht aus vier Kriterien, welche die unterschiedlichen Ausprägungen der Informationsverknüpfung beschreiben. Hierbei handelt es sich um die Referenzdimension, die Art der räumlichen Verknüpfung, die visuelle Kontinuität sowie das Vorhandensein von Kontext. Diese werden genutzt, um in der Literatur beschriebene Darstellungsvarianten zu klassifizieren. Hierdurch wird zum einen die Anwendbarkeit der Kriterien geprüft und zum anderen eine Basis für weitere Untersuchungen geschaffen.

Schlüsselwörter

Augmented Reality, Information Presentation, Interface Design

CR-Kategorien

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Artificial, augmented, and virtual realities*

1 Einleitung

Augmented-Reality-Anwendungen finden in den letzten Jahren, durch Fortschritte in Soft- und Hardware, auch außerhalb von Forschungsprojekten eine zunehmende Verbreitung [19]. Für die Industrie stellt die Unterstützung bei Wartungs- und Reparaturarbeiten durch Augmented Reality ein interessantes und viel beachtetes Anwendungsfeld dar. Unter Augmented Reality (kurz AR) wird nach Azuma das Einblenden von virtuellen Informationen in das Sichtfeld eines Nutzers verstanden, welche in der physischen Welt registriert sind und auf Änderungen des Sichtbereichs in Echtzeit reagieren [1]. Der Begriff Registrierung, beziehungsweise geometrische Registrierung, beschreibt das korrekte Einpassen von virtuellen Artefakten in

---

*

Betreuer Hochschule: Prof. Dr.-Ing. Cristóbal Curio
Hochschule Reutlingen
Cristobal.Curio@Reutlingen-University.de
Betreuer Firma: Tobias Müller
Robert Bosch GmbH
Tobias.Mueller8@bosch.com

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

die physische Welt [3].
Studien belegen, dass eine Unterstützung durch AR bei prozeduralen Aufgaben, wie Wartungs- und Reparaturarbeiten, zu einer geringeren Fehlerhäufigkeit und einer schnelleren Durchführung der Aufgabe führen [9][20]. Prozedurale Aufgaben lassen sich grundsätzlich in informations- (kognitive) sowie arbeitsbezogene (psychomotorische) Aktivitäten untergliedern [14]. Durch den Einsatz eines AR-Systems wird ein Wechsel zwischen beiden Aufgaben (engl. attention switching) minimiert, welches die mentale Belastung des Nutzers verringert [20].

1.1 Herausforderungen

Damit der mentale Aufwand so gering wie möglich gehalten werden kann, müssen neben rein technischen Herausforderungen (Tracking und Registrierung), der allgemeinen Interaktion mit dem System, insbesondere auch die Art und Weise wie Informationen dargestellt und eingeblendet werden berücksichtigt werden. Eine Vielzahl von wahrnehmungsbezogenen Problematiken lassen sich auf technische Limitierungen zurückführen [11]. Ein großer Anteil wird jedoch durch fehlendes Verständnis und inadäquate Methoden zur Darstellung von Informationen verursacht. Für AR-Anwendungen existieren, anders als für klassische Desktop oder Webanwendungen, kaum Richtlinien zur Gestaltung der Benutzerschnittstelle [13]. In den letzten Jahren erfahren nach [26] Visualisierungsthemen im Allgemeinen sowie die Definition von Design Richtlinien und Taxonomien [13][18][21][24] eine zunehmende Beachtung. Das Visualisieren von semantischen Verknüpfungen zwischen physischen und virtuellen Objekten, welches als Informationsverknüpfung bezeichnet wird, stellt hierbei ein wichtiger Aspekt dar. Ein naives Einblenden und Überlagern von Informationen unmittelbar über dem zugehörigen physischen Referenzobjekt führt unausweichlich zu Verdeckungen. Dies beeinträchtigt den Nutzer mehr als es ihm Nutzen bringt [13].

1.2 Zielsetzung

Im Rahmen dieser Ausarbeitung wird ein Klassifikationsschema entwickelt , mit dem sich Darstellungsvarianten hinsichtlich der Informationsverknüpfung kategorisieren lassen. Hierfür werden zunächst die betrachteten Komponenten identifiziert und definiert, bevor anschließend Kriterien aufgestellt werden, mit denen sich unterschiedliche Ausprägungen beschreiben lassen. Durch eine Klassifizierung von Darstellungsvarianten, welche in der Literatur für den prozeduralen Anwendungsfall beschrieben werden, wird eine Übersicht geschaffen und das Klassifikationsschema auf Anwendbarkeit geprüft. Eine Berücksichtigung von Interaktionsschnittstellen sowie von haptischen oder akustischen Informationen und weiteren Formen von AR findet nicht statt. Es ist jedoch denkbar, dass die hier aufgestellten Kategorien auch außerhalb dieses Anwendungsfalls gültig sind.

2 Verwandte Arbeiten

Durch die Möglichkeiten von Augmented Reality, virtuelle Objekte in die physische Welt einzublenden, wird eine räumliche als auch zeitliche Beziehung (engl. spatio-temporal relationship) zwischen der physischen Umgebung und den virtuellen Inhalten geschaffen [22]. Dieser komplexe Darstellungsraum übersteigt die Möglichkeiten von klassischen WIMP-basierten Benutzerschnittstellen [21]. Tönnis et al. unterteilt diesen in die physische sowie virtuelle Welt und führt eine sogenannte Referenzdimension ein, welche virtuelle und physische Objekte miteinander verknüpft [21]. Des Weiteren werden fünf Kategorien definiert, welche eine allgemeine Klassifizierung von Varianten im Darstellungsraum ermöglichen. Müller unterteilt den Darstellungsraum, in Abhängigkeit vom räumlichen Bezug der dargestellten Informationen, in fünf Ebenen [13]. Die unteren zwei Ebenen beinhalten Informationen der physischen Welt, welche direkt oder indirekt (zum Beispiel durch ein Video-See-Through-System) wahrgenommen werden. Ebene drei und vier umfassen virtuelle Informationen, welche räumlich referenziert und auf dritter Ebene von räumlicher Natur sein müssen. Auf oberster Ebene befinden sich Informationen, die keine Verbindung zur physischen Welt aufweisen [13]. Wither et al. präsentieren ein detailliertes Klassifikationsschema für den Anwendungsfall von Annotationen [25]. In [17] werden vier Darstellungsvarianten mittels einer Nutzerstudie hinsichtlich dem Einfluss von Kontextobjekten untersucht. Lediglich zwei Varianten sind jedoch als AR im Sinne von Azuma zu verstehen [1]. Vincent et al. definiert drei Ebenen, die miteinander räumlich verknüpft (engl. spatial mapping) sind [22]. Die genannten Ausarbeitungen stellen eine gute konzeptionelle Basis für die Betrachtung der Informationsverknüpfung dar. In keiner der genannten Arbeiten wird die Informationsverknüpfung im Detail untersucht, Teile können jedoch genutzt werden um Komponenten zu identifizieren sowie Kriterien für das Klassfikationsschema abzuleiten.

3 Informationsverknüpfung

Das Einblenden von räumlich registrierten virtuellen Objekten in das Sichtfeld des Nutzers stellt die Kernaufgabe eines jeden AR-Systems dar. Das Darstellen von semantischen Verknüpfungen zwischen virtuellen und physischen Objekten, welche durch ein AR-Interface betrachtet werden, wird im Kontext dieser Ausarbeitung als Informationsverknüpfung bezeichnet. Für eine genauere Betrachtung der Informationen im Darstellungsraum sowie ihren Beziehungen untereinander, müssen die einzelnen Komponenten zunächst identifiziert werden. Grundsätzlich muss jede eingeblendete Information einen physischen Bezugspunkt im Sichtbereich haben. Die virtuellen Informationen können in unterschiedliche Bestandteile aufgeschlüsselt werden. Wither et al. unterteilen beispielsweise Annotationen in räumlich abhängige und unabhängige Komponenten [25]. Diese Unterteilung besitzt jedoch nur für diesen Anwendungsfall Gültigkeit. Für eine Generalisierung lassen sich virtuelle Informationen in informationstragende sowie informationsverknüpfende Bestandteile aufteilen. Zusätzlich können Informationen eingeblendet werden, welche keine räumlichen Bezug besitzen (Metaobjekte). Eine Unterteilung dieser Art existiert nach bestem Wissen des Autors in der Literatur bisher nicht. Die identifizierten Komponenten werden in den folgenden Abschnitten definiert und sind in Abbildung 1 veranschaulicht.

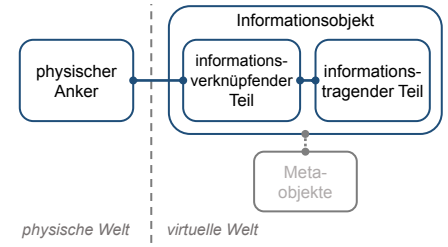

Abbildung 1: Komponenten und ihre Beziehungen im Darstellungsraum

3.1 Physischer Anker

Hierbei handelt es sich um ein physisches Objekt oder einen Bereich von Interesse (engl. Point-of-Interest), welches als Bezugspunkt für eine Augmentierung dient. Dieser kann unterschiedliche Ausprägungen haben und wird beispielsweise von Wither et al. als *location complexity* beschrieben [25]. Ein physischer Anker kann im einfachsten Fall eine 3D Koordinate (3 DOF) im Weltkoordinatensystem sein. Wird diese um die Orientierung ergänzt (6 DOF), können Objekte räumlich auf Oberflächen ausgerichtet werden. Komplexere Strukturen können durch mehrere Koordinaten als Linie, 2D Boundingbox oder als komplexe 3D Modelle beschrieben werden [25]. Für eine differenzierte Unterscheidung wird diese Ausprägung im Folgenden als *räumliche Gestalt* bezeichnet und lässt sich in dieser Form in der Literatur bisher nicht finden.

3.2 Informationsobjekt

Das Informationsobjekt stellt eine sinntragende Augmentierung dar, welche Informa-

tionen, in beliebiger Form, zu einem physischem Objekt bereitstellt. Als sinntragende Augmentierung wird im Kontext einer prozeduralen Aufgabe eine Beschreibung einer einzelnen Handlung oder eines Arbeitsschritts verstanden. Diese hat zu mindestens einem physischen Anker im aktuellen Sichtbereich eine Relation und steht zu diesem in räumlichen Bezug. Ein Informationsobjekt lässt sich in einen informationsverknüpfenden sowie einen informationstragenden Teil untergliedern. Beide Bestandteile können hierbei visuell getrennt oder als Einheit, wie beispielsweise in Form einer Sprechblase, wahrgenommen werden. Sind diese visuell getrennt, müssen beide Komponenten ebenfalls zueinander in Bezug gebracht werden, beispielsweise in Form von alphanumerischen Zeichen oder Verbindungslinien.

- **Informationsverknüpfender Teil:** Wie bereits definiert, muss jedes Informationsobjekt räumlich zu einem physischen Objekt in Bezug stehen. Diese visuelle Beziehung wird durch den informationsverknüpfenden Teil geschaffen beziehungsweise beschrieben. Dies kann auf unterschiedliche Arten, wie direkter Überlagerung (3D Modell), räumlicher Nähe (Gestaltheuristiken) oder in Form von Pfeilen, hergestellt werden und liegt somit im Fokus dieser Ausarbeitung. Der informationsverknüpfende Teil stellt hierbei nicht immer ein explizit visuell identifizierbares Objekt, wie beispielsweise eine Verbindungslinie dar, sondern kann sich zum Beispiel bei der Verwendung von Gestaltheuristiken lediglich auf die räumliche Positionierung der informationstragenden Komponente auswirken.
- **Informationstragender Teil:** Der informationstragende Teil ist eine semantische Information über den physischen Bezugspunkt und stellt eine Arbeitsanweisung dar. Im Kontext von prozeduralen Aufgaben handelt es sich in der Regel um textuelle Anweisungen, Illustrationen (Explosionszeichnungen) oder 3D Animationen von Werkzeugen [7].

Wird das Informationsobjekt betrachtet, kann dieses unterschiedliche räumliche Ausprägungen aufweisen und ist von den verfügbaren Daten beziehungsweise dem Anwendungsfall abhängig. Wither et al. bezeichnet dies als *content complexity* und unterscheidet zwischen einfachen 2D Text Annotationen bis hin zu komplex animierten 3D Modellen [25]. Tönnis et al. beschreibt dies als *dimensionality of represented features* und bezieht sich auf die Abbildung und Integration virtueller Inhalte in die physische Umgebung und unterteilt diese ebenfalls in 2D und 3D Objekte [21]. In dieser Ausarbeitung wird ein anderer Ansatz verwendet, welcher die Informationsobjekte anhand ihrer räumlichen Natur unterscheidet. Ein Objekt ist von räumlicher Natur, wenn sich dieses in die physische Umgebung geometrisch korrekt integrieren lässt. Diese Unterscheidung wird in [13] zur Differenzierung zwischen der dritten und vierten Ebene genutzt. Beispiele für Objekte, welche keine räumliche Natur besitzen, stellen 2D Text-Annotationen aber auch 3D Modelle dar, die lediglich zur Veranschaulichung dienen und sich räumlich nicht in den aktuellen Kontext integrieren lassen. Das Einblenden eines CAD-Modells, welches ein physisches Bauteil überlagern kann, stellt ein Beispiel für ein Objekt räumlicher Natur dar. Letzteres bietet für den Betrachter ein natürlicheres Erscheinungsbild dar und führt zu einer konsistenten Umgebung (engl. consistent environment) [21].

3.3 *Metaobjekte*

Metaobjekte stellen virtuelle Objekte dar, die zusätzliche Informationen bereitstellen und keinen physischen Anker besitzen. Beispiele hierfür sind zusätzliche Illustrationen oder Fortschrittsanzeigen. Metaobjekte werden im Folgenden nicht näher betrachtet, da sie nach Definition von Azuma auch kein AR darstellen [1].

4 Klassifikationsschema

Für eine genaue Betrachtung und Differenzierung von bestehenden Varianten zur Informationsverknüpfung ist es notwendig Kriterien aufzustellen, welche die einzelnen Ausprägungen beschreiben und einen Vergleich zulassen. Die zuvor beschriebenen Komponenten bilden für das Identifizieren von Kriterien eine Basis. Betrachtet man den informationsverknüpfenden Teil, kann hinsichtlich der räumlichen Nähe zum physischem Anker und die Art und Weise, wie diese geschaffen wird, unterschieden werden (vgl. visuelle Kontinuität, Art der räumlichen Verknüpfung). Des Weiteren können Bestandteile des Informationsobjekts in verschiedenen Darstellungsräumen (vgl. Referenzdimension) platziert, beziehungsweise verankert werden. Den physischen Anker umgebende Objekte können hervorgehoben werden, um Informationen räumlich besser einzuordnen (vgl. Kontext). Wie bereits in Kapitel 2 aufgeführt, werden unter anderem in [4], [21], [22], sowie [25] Taxonomien beschrieben, die hier ebenfalls berücksichtigt werden. Insbesondere das Klassifikationsschema von Tönnis et al., das aus fünf Dimensionen besteht und eine Klassifizierung von Augmentierungen hinsichtlich ihrer Ausprägungen im Darstellungsraum ermöglicht, stellt eine gute Basis dar [21]. Bezogen auf die Informationsverknüpfung lassen sich die Dimensionen *Mounting Point* und *Type of Reference* verwenden und werden im weiteren Verlauf genauer betrachtet. Die Dimensionen *Temporalität*, die Unterscheidung von zeitlich diskreten und kontinuierlichen Visualisierungen, sowie dem *Frame of Reference* (abhängig vom Endgerät, egozentrischer vs. exozentrischer Viewpoint), werden nicht weiter betrachtet. Im Folgenden werden vier Kriterien vorgestellt, welche die Ausprägungen der Informationsverknüpfung beschreiben [1].

[1] Die Definitionen der Kriterien wurden in gemeinschaftlicher Arbeit zusammen mit dem Betreuer Tobias Müller aufgestellt.

4.1 *Referenzdimension*

Grundsätzlich können Informationsobjekte auf verschiedene räumliche Dimensionen beziehungsweise Koordinatensysteme aufgeteilt werden. Positionen und Orientierungen in der physischen Welt werden in Weltkoordinaten (kurz WKS) beschrieben. Stellen die Augen, beziehungsweise die Kamera, den Ursprung dar, wird von einem Betrachterkoordinatensystem (kurz BKS) beziehungsweise Kamerakoordinatensystem (kurz KKS) gesprochen. Bei der Verwendung von Head-Mounted-Displays entspricht das BKS dem KKS und wird als egozentrische Betrachtungsweise bezeichnet [21]. Um die Betrachtung unabhängig vom verwendeten Endgerät zu halten, wird hier nur das WKS sowie das KKS betrachtet. Die Ausprägungen werden anhand folgender Kategorien unterschieden:

- **WKS**: Das Informationsobjekt befindet sich ausschließlich im Weltkoordinatensystem und es existieren keinerlei Abhängigkeiten zum Betrachterkoordinatensystem.
- **KKS-orientiert**: Die Position des Informationsobjekts befindet sich im WKS. Die Orientierung, beziehungsweise die Ausrichtung des Informationsobjekts, ist dem Betrachter zugewandt.
- **KKS-ergänzt**: Ein Objekt räumlicher Natur, welches im WKS positioniert und orientiert ist, wird durch Informationen im KKS ergänzt.
- **KKS-positioniert**: Das Informationsobjekt ist auf KKS und WKS verteilt. Die Position des informationstragenden Teils befindet sich im KKS. Die Orientierung ist jedoch abhängig vom physischen Anker (WKS).
- **KKS-verortet**: Das Informationsobjekt ist auf KKS und WKS verteilt. Der informationstragende Teil befindet sich im KKS und der informati-

onsverknüpfende Teil stellt die Relation zwischen KKS und WKS her.

- **KKS**: Objekte, welche ausschließlich im KKS positioniert und orientiert sind, stellen kein AR dar [1] und werden nicht weiter betrachtet.

Eine ähnliche Unterscheidung hinsichtlich der Referenzdimension und geometrischen Eigenschaften wird auch von Tönnis et al. beschrieben und als *Mounting Dimension* bezeichnet [21]. Es wird zwischen *Human*, *Environment*, *World* sowie *Multiple Mountings* unterschieden. Die oben definierten Ausprägungen stellen insbesondere eine Spezialisierung des letzten Anwendungsfalls dar und ermöglichen hierdurch eine spezifischer Beschreibung hinsichtlich der Informationsverknüpfung.

4.2 Art der räumlichen Verknüpfung

Diese Kategorie beschreibt die geometrische beziehungsweise räumliche Verknüpfung (engl. spatial mapping) zwischen physischem Anker und Informationsobjekt. Diese können prinzipiell in direkte und indirekte Verknüpfungen unterteilt werden [24]. Als direkt wird nach Webel et al. das Einblenden von Objekten räumlicher Natur (3D Modelle) welche den physischen Anker komplett überlagern oder zu diesem räumlich korrekt in Bezug stehen verstanden. Als indirekte Visualisierungen werden räumlich distanzierte Informationen in Form von Annotationen bezeichnet [24]. Webel et al. nutzen diese Unterscheidung, um die Auswirkung auf das Lernverhalten des Nutzers bei der Durchführung von prozeduralen Aufgaben zu untersuchen. Tönnis et al. unterscheidet zwischen direkten, indirekten und rein referenziellen Darstellungsformen (vgl. *type of reference*) [21]. Diese Betrachtungsweise bezieht sich jedoch allein auf die Sichtbarkeit des physischen Ankers und eignet sich für die Betrachtung der Informationsverknüpfung weniger. In [22] wird die räumliche Verknüpfung in *konformal*, *partiell*, *distanzierte* und *außerhalb* des Sichtbereichs (Off-Screen) unterschieden. Die Unterscheidung in konformal sowie distanziert entspricht der Differenzierung in direkt und indirekt von Webel et al. Insbesondere das Verständnis von partiell bezieht sich nach Wither et al. jedoch mehr auf die Referenzdimension. Dies wird in der vorliegenden Ausarbeitung als gesondertes Kriterium aufgeführt und genauer definiert (vgl. 4.1 Referenzdimension).
Da in der Literatur keine geeignete Aufteilung gefunden wurde, wird in dieser Ausarbeitung die Art der räumlichen Verknüpfungen in drei Ausprägungen unterschieden und wie folgt definiert:

- **Direkt**: Als direkt wird das Einblenden von Objekten räumlicher Natur (3D Modelle), welche hinsichtlich räumlicher Position und Ausrichtung mit dem physischen Anker übereinstimmen oder zu diesem räumlich korrekt in Bezug stehen (eingeblendete Werkzeuge), bezeichnet. Hierbei können Ist-, Zwischen-, oder Sollzustände inklusive Übergänge beschrieben werden.

- **Partiell**: Ist der informationstragende Teil des Informationsobjektes visuell wahrnehmbar auf direkte und indirekte Verknüpfungen aufgeteilt, wird dies als partielle Verknüpfung angesehen.

- **Indirekt**: Hierbei ist der informationstragende Teil des Informationsobjektes vom physischen Anker räumlich getrennt, ohne das diese Darstellung einen Ist-, Zwischen- oder Sollzustand beschreibt. Der informationsverknüpfende Teil stellt die Verbindung her.

4.3 Visuelle Kontinuität

Die visuelle Wahrnehmung des Menschen lässt sich aufgrund der unterschiedlichen Rezeptorendichte des Auges in die drei Bereiche, foveal, parafoveal sowie peripher, unterteilen [12]. In dem relativ kleinen kegelförmigen fovealen Teil des Sichtbereichs (fovea centralis) ist die Sehschärfe am höchsten. Eine Unterscheidung in die Sichtbereiche lässt sich nutzen, um eine Aussage über die räumliche Nähe von physischen zu virtuellen Objekten zu treffen. Aufgrund der Abhängigkeit vom Abstand des Betrachters zum betrachteten Objekt auf die Sehschärfe, wird die Grundannahme getroffen, dass sich der Abstand zum physischen Objekt im Interaktionsradius des Benutzers befindet. Für eine Betrachtung werden folgende Unterscheidungen definiert und finden sich nach bestem Wissen des Autors in keiner bisherigen wissenschaflichen Ausarbeitung wieder:

- **Räumliche Nähe**: Die Zusammengehörigkeit wird durch räumliche Nähe dargestellt. Hierbei stellt der informationsverknüpfende Teil ein Objekt räumlicher Natur dar, welches den physischen Anker überlagert oder mit einem räumlichen Versatz (Offset) zu diesem dargestellt wird. Der informationsverknüpfende Teil befindet sich im fovealen oder parafovealen Sichtbereich und kann gleichzeitig auch informationstragende Bestandteile beinhalten.

- **Kontinuierlich**: Der Abstand zwischen dem informationsverknüpfenden Teil und physischem Anker befindet sich im fovealen Bereich. Der informationsverknüpfende Teil stellt eine visuell durchgehende Verbindung zum informationstragenden Teil her.

- **Symbolisch**: Als symbolisch wird eine Verknüpfung dann bezeichnet, wenn sich ein Teil des informationsverknüpfenden Teils in räumlicher Nähe zum physischen Anker und der andere in der Nähe zum informationstragenden Teil befindet. Zwischen diesen besteht keine visuell durchgehende Verbindung. Die Verknüpfung wird symbolisch, zum Beispiel durch entsprechende Farbgebung, Symbole, alphanumerische Zeichen oder anderen Hilfsmitteln hergestellt. Hierdurch

Abbildung 2: Verwendung von symbolischen Verknüpfung [9].

können Teile des Informationsobjektes auch im parafovealen Bereich dargestellt werden (vgl. Abbildung 2).

- **Diskontinuierlich**: Existiert keine visuelle Verbindung zu einem physischen Objekt wird dies als diskontinuierlich bezeichnet. Dies ist nach Azuma [1] auch keine AR und wird nicht weiter betrachtet.

4.4 Kontext

Werden virtuelle Objekte in die physische Umgebung eingeblendet, kann dies zu Ambiguitäten bei der Tiefenwahrnehmung führen. Eine Möglichkeit die Tiefenwahrnehmung zu verbessern, stellt das Einblenden von sogenannten Kontextobjekten dar [10][17]. Hierbei werden Hilfslinien oder Konturen von physischen Objekten, welche eine Relation zum physischen Anker besitzen, virtuell hervorgehoben (vgl. Focus and Context [10]). Es ist naheliegend, dass sich Probleme bei der Tiefenwahrnehmung unmittelbar auf die Informationsverknüpfung auswirken. Daher ist es sinnvoll Kontextobjekte einzublenden und ist prinzipiell unabhängig von der verwendeten Darstellungsvariante. Darauf aufbauend werden Darstellungsvarianten danach unterschieden ob sie eine Kontextvisualisierung nutzen oder nicht.

Tabelle 1: Übersicht der definierten Kriterien und ihre Ausprägungen

Kriterien	Eigenschaften
ADV	- direkt - partiell - indirekt
RD	- WKS - KKS-orientiert - KKS-ergänzt - KKS-positioniert - KKS-verortet - KKS[2]
VK	- kontinuierlich - räumliche Nähe - symbolisch - diskontinuierlich[2]
K	- ja - nein

Tabelle 2: Übersicht der definierten Dateneigenschaften

Kriterien	Eigenschaften
DA	- 3-DOF - 6-DOF - räumliche Gestalt
DI	- nicht räumlicher Natur - räumliche Natur

5 Anwendung

In diesem Abschnitt wird das definierte Klassifikationsschema angewendet, indem unterschiedliche Darstellungsvarianten kategorisiert werden. Hierdurch sollen zum einen die vier Kategorien auf Anwendbarkeit geprüft werden und zum anderen soll hierdurch eine Übersicht über Darstellungsvarianten gegeben werden. Für eine bessere Lesbarkeit werden die Bezeichnungen der vier definierten Kriterien zunächst wie folgt abgekürzt; Art der räumlichen Verknüpfung (kurz *ADV*), Referenzdimension (kurz *RD*), visuelle Kontinuität (kurz *VK*), Kontext (*K*). In Tabelle 1 sind die Kriterien und ihre Ausprägungen zusammenfassend aufgelistet. Die Dateneigenschaften der Komponenten; Dimension des physischen Ankers (kurz *DA*) sowie Dimension des Informationsobjekts (kurz *DI*) finden sich in Tabelle 2 wieder.

5.1 Klassifizierung

In der Literatur werden eine Vielzahl von Darstellungsvarianten, für die Unterstützung bei Wartungs- und Reparaturarbeiten, beschrieben. Um die Anwendbarkeit des Klassifikationsschemas zu prüfen, wurde daher eine möglichst große Anzahl von einschlägiger Literatur berücksichtigt. Eine Auflistung und Klassifikation der betrachteten Varianten findet sich in Tabelle 3 wieder. Ähnliche Varianten wurden vorab gefiltert und es wurde darauf geachtet möglichst unterschiedliche Darstellungsformen zu berücksichtigen.
Im Folgenden wird exemplarisch das Einblenden eines Bewegungsablaufs klassifiziert. Hier kann ein zu verwendendes Werkzeug, wie zum Beispiel ein Schraubenzieher, oder eine spezifische Handbewegung eingeblendet werden, um einen durchzuführenden Arbeitsschritt zu visualisieren [9][16]. In Abbildung 3 ist ein animierter Bewegungsablauf dargestellt, der dem Nutzer zeigt wie ein Bauteil aus einer Autotür zu entfernen ist. Der Bewegungsablauf sowie die eingeblendete Hand stellen hierbei den informationstragenden Teil dar. Das farblich hervorgehobene CAD-Model sowie die Hand stellen informationsverknüpfende Teile dar. Das Informationsobjekt ist von räumlicher Natur, da das CAD-Modell als auch die Hand in Größe und Form geometrisch korrekt in die physische Umgebung integriert sind. Zusätzlich stehen Position und Ausrichtung mit dem physischem Anker räumlich korrekt in Bezug. Durch den Bewegungsablauf wird ein Übergang vom Ist- zum Sollzustand beschrieben. Daher ist die Art der räumlichen Verknüpfung direkt. Die Ausprägung der visuellen Kontinuität ist räumliche Nähe, da das CAD-Modell den physischen Anker überlagert und auch die Hand in räumlicher Nähe zum physischen Anker ist. Die eingeblendeten Informationen befinden sich ausschließlich im WKS. Eine Kontextvisualisierung wird in diesem Anwendungsfall nicht genutzt.

[2]Wird nicht betrachtet, da es sich nach [1] nicht um AR handelt.

Tabelle 3: Übersicht von klassifizierten Darstellungsvarianten (rN = räumliche Natur)

Informationsobjekt	*ADV*	*RD*	*VK*	*K*	*DA*	*DI*
Text auf Oberfläche [6]	direkt	WKS	räumliche Nähe	nein	6DOF	rN
Bewegungsablauf [7] [16]	direkt	WKS	räumliche Nähe	nein	räumlich	rN
Darstellung verdeckter Objekte [4]	partiell	WKS	räumliche Nähe	ja	räumlich	rN
3D-Modell mit Annotation [8][23]	partiell	KKS-ergänzt	räumliche Nähe	nein	räumlich	rN
3D Modell in räumlicher Nähe [17]	partiell	WKS	räumliche Nähe	ja	räumlich	rN
Annotation mit Verbindungslinie [15]	indirekt	KKS-verortet	kontinuierlich	nein	3DOF	nicht rN
Vergrößerte Ansicht des POI [5]	indirekt	KKS-positioniert	symbolisch	nein	räumlich	rN
benutzerzentrierte Beschriftungen [2]	indirekt	KKS-orientiert	kontinuierlich	nein	3DOF	nicht rN
symbolische Annotation [9][23]	indirekt	KKS-verortet	symbolisch	nein	3DOF	nicht rN

5.2 Diskussion

Wie sich bei der Klassifizierung gezeigt hat, konnten alle Varianten anhand der vier definierten Kategorien beschrieben werden. Aus den Kategorien lassen sich verschiedene Hypothesen ableiten, die Anhaltspunkte für weitere Untersuchung bieten. So ist anzunehmen, dass der mentale Aufwand, eine Verbindung zum physischen Anker herzustellen, bei der Verwendung von räumlich direkten Verknüpfungen geringer ist. Bezüglich der Referenzdimension lässt sich annehmen, dass der mentale Aufwand bei WKS sowie KKS-orientierten Darstellungsvarianten geringer ist. Betrachtet man die visuelle Kontinuität ist anzunehmen, dass räumliche Nähe und visuell kontinuierliche Varianten symbolischen Verknüpfungen vorzuziehen sind. Ausprägungen der geometrischen Dimensionalität (*DA* / *DI*) werden als gegeben betrachtet und sind vom Anwendungsfall abhängig. Für eine Untersuchung der aufgestellten Hypothesen können durch das Klassifikationsschema Darstellungsvarianten voneinander abgegrenzt und ähnliche

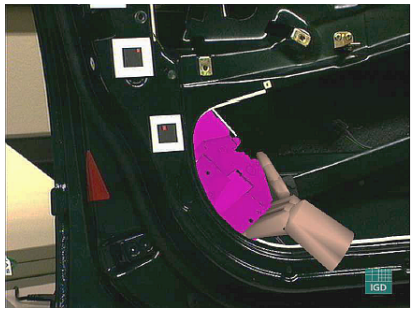

Abbildung 3: Direkte Verknüpfung durch räumliche Nähe im WKS [16].

Varianten aggregiert werden.
Unabhängig von der verwendeten Darstellungsvariante besteht grundsätzlich die Gefahr, dass physische Objekte oder andere relevante Arbeitsbereiche durch virtuelle Objekte verdeckt werden können. Diese Verdeckungsproblematik wird auch als *visual interference* beschrieben [20]. Insbesondere bei der Verwendung von direkten und partiellen Verknüpfungen lassen sich Verdeckungen nicht vermeiden. Erst durch eine Verwendung von indirekten Verknüpfungen entstehen Möglichkeiten, wie beispielsweise durch symbolische Verknüpfungen oder Verbindungslinien, Verdeckungen zu umgehen. Der beschriebene Sachverhalt steht mit der zuvor getroffenen Annahme, räumlich direkte Verknüpfungen bedeuten einen geringeren mentalen Aufwand, im Konflikt und muss weiter untersucht werden.

5.3 Limitierungen

Ein Aspekt, welcher in der Literatur bisher kaum betrachtet wurde, stellen komplexere Animation dar, welche über das Einblenden eines animierten Werkzeugs hinausgehen (vgl. MARTA Projekt von VW [19]). Durch komplexere Animationen können fließende Übergänge zwischen den definierten Kategorien geschaffen werden. Ein CAD-Modell könnte beispielsweise zunächst räumlich direkt dargestellt werden, bevor es anschließend als Explosionszeichnung an einer festen Position im Kamerakoordinatensystem dargestellt wird. Dies würde neben einer Veränderung der Referenzdimension (WKS zu KKS-orientiert) auch alle weiteren Ausprägungen beeinflussen. Grundsätzlich kann daher zwischen einfachen Animationen, welche zu keiner Änderung der aufgestellten Ausprägungen führen, und komplexen Animationen unterschieden werden. Erstere umfasst die Beschreibung von direkten räumlichen Verknüpfungen indem Ist-, Zwischen, und Sollzuständen und ihre Übergänge einbezogen werden. Für eine Betrachtung von komplexeren Animationen stößt das Klassifikationsschema jedoch an seine Grenzen und erfordert eine gesonderte Betrachtungsweise.

6 Zusammenfassung & Ausblick

In dieser Ausarbeitung wurde ein neuartiges Klassifikationsschema vorgestellt, das eine Betrachtung und Klassifikation von Darstellungsvarianten hinsichtlich ihrer Informationsverknüpfung ermöglicht. Neben der Identifikation und Definition der betrachteten Komponenten, physischer Anker sowie des Informationsobjektes, wurden die Kategorien Referenzdimension, Art der räumlichen Verknüpfung, Visuelle Kontinuität sowie Kontext definiert. Die aufgestellten Komponenten sowie die Kategorien wurden in dieser Form in der Literatur bisher nicht beschrieben. Hierdurch können die unterschiedlichen Ausprägungen der Informationsverknüpfung beschreiben werden und ermöglichen eine Klassifizierung von Darstellungsvarianten. Dies führt zu einem besseren Verständnis über die verschiedenen Ansätze und lässt sich nutzen um Darstellungsvarianten voneinander abzugrenzen. Zusätzlich können durch Kombination der einzelnen Ausprägungen Varianten konstruiert werden, welche bisher nicht berücksichtigt wurden. Diese Arbeit wird als Basis für weitere Untersuchungen über die Auswirkungen der einzelnen Ausprägungen auf den mentalen und sensorischen Aufwand genutzt.

Literatur

[1] R. T. Azuma et al. A survey of augmented reality. *Presence*, 6(4):355–385, 1997.

[2] B. Bell, S. Feiner, and T. Höllerer. View management for virtual and augmented reality. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pages 101–110. ACM, 2001.

[3] R. Dörner, W. Broll, P. Grimm, and B. Jung. *Virtual und augmented reality (VR/AR): Grundlagen und Methoden der Virtuellen und Augmentierten Realität*. Springer-Verlag, 2014.

[4] N. Elmqvist and P. Tsigas. A taxonomy of 3d occlusion management for visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 14(5):1095–1109, 2008.

[5] T. Engelke, S. Webel, and N. Gavish. Generating vision based lego augmented reality training and evaluation systems. In *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on*, pages 223–224. IEEE, 2010.

[6] T. Götzelmann, K. Hartmann, and T. Strothotte. Annotation of animated 3d objects. In *SimVis*, volume 7, pages 209–222. Citeseer, 2007.

[7] S. Henderson and S. Feiner. Exploring the benefits of augmented reality documentation for maintenance and repair. *Visualization and Computer Graphics, IEEE Transactions on*, 17(10):1355–1368, 2011.

[8] S. J. Henderson and S. Feiner. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, pages 135–144. IEEE, 2009.

[9] S. J. Henderson and S. K. Feiner. Augmented reality in the psychomotor phase of a procedural task. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pages 191–200. IEEE, 2011.

[10] D. Kalkofen, E. Mendez, and D. Schmalstieg. Interactive focus and context visualization for augmented reality. In *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 1–10. IEEE Computer Society, 2007.

[11] E. Kruijff, J. E. Swan II, and S. Feiner. Perceptual issues in augmented reality revisited. In *ISMAR*, volume 9, pages 3–12, 2010.

[12] L. M. Lorenz. *Entwicklung und Bewertung aufmerksamkeitslenkender Warn- und Informationskonzepte für Fahrerassistenzsysteme: Aufmerksamkeitssteuerung in der frühen Phase kritischer Verkehrssituationen*. PhD thesis, Universitätsbibliothek der TU München, 2014.

[13] T. Müller. Towards a framework for information presentation in augmented reality for the support of procedural tasks. In *Augmented and Virtual Reality*, pages 490–497. Springer, 2015.

[14] U. Neumann and A. Majoros. Cognitive, performance, and systems issues for augmented reality applications in manufacturing and maintenance. In *Virtual Reality Annual International Symposium, 1998. Proceedings., IEEE 1998*, pages 4–11. IEEE, 1998.

[15] J. Platonov, H. Heibel, P. Meier, and B. Grollmann. A mobile markerless ar system for maintenance and repair. In *Proceedings of the 5th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 105–108. IEEE Computer Society, 2006.

[16] D. Reiners, D. Stricker, G. Klinker, and S. Müller. Augmented reality for construction tasks: Doorlock assembly. *Proc. IEEE and ACM IWAR*, 98(1):31–46, 1998.

[17] C. M. Robertson, B. MacIntyre, and B. N. Walker. An evaluation of graphical context when the graphics are outside of the task area. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 73–76. IEEE Computer Society, 2008.

[18] C. Rolim, D. Schmalstieg, D. Kalkofen, and V. Teichrieb. Design guidelines for generating augmented reality instructions. In *In Proc. International Symposium on Mixed and Augmented Reality (ISMAR 2015) Posters*, 2015.

[19] D. Stanimirovic, N. Damasky, S. Webel, D. Koriath, A. Spillner, and D. Kurz. [poster] a mobile augmented reality system to assist auto mechanics. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, pages 305–306. IEEE, 2014.

[20] A. Tang, C. Owen, F. Biocca, and W. Mou. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 73–80. ACM, 2003.

[21] M. Tönnis, D. A. Plecher, and G. Klinker. Representing information–classifying the augmented reality presentation space. *Computers & Graphics*, 37(8):997–1011, 2013.

[22] T. Vincent, L. Nigay, and T. Kurata. Classifying handheld augmented reality: Three categories linked by spatial mappings. In *Workshop on Classifying the AR Presentation Space at ISMAR 2012*, 2012.

[23] S. Webel, U. Bockholt, T. Engelke, N. Gavish, M. Olbrich, and C. Preusche. An augmented reality training platform for assembly and maintenance skills. *Robotics and autonomous systems*, 61(4):398–403, 2013.

[24] S. Webel, U. Bockholt, T. Engelke, N. Gavish, and F. Tecchia. Design recommendations for augmented reality based training of maintenance skills. In *Recent Trends of Mobile Collaborative Augmented Reality Systems*, pages 69–82. Springer, 2011.

[25] J. Wither, S. DiVerdi, and T. Höllerer. Annotation in outdoor augmented reality. *Computers & Graphics*, 33(6):679–689, 2009.

[26] F. Zhou, H. B.-L. Duh, and M. Billinghurst. Trends in augmented reality tracking, interaction and display: A review of ten years of ismar. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 193–202. IEEE Computer Society, 2008.

Komplexe Regeln in einer Internet-of-Things-Anwendung *

Peter Einberger
Reutlingen University
Peter.Einberger@Student.Reutlingen-University.de

Abstract

Diese Arbeit behandelt das Integrieren von komplexen Regeln in die Industrie 4.0 Internet-of-Things-Anwendung Sense&Act, die am Fraunhofer Institut für Produktionstechnik und Automatisierung IPA entwickelt wird. Sensoren mit denen Maschinen nachgerüstet werden, senden Werte über bestehende Internet Infrastruktur an einen Regelserver, der auf Basis von Nutzer definierten Regeln Aktionen auslöst. Um Regeln anhand mehrerer Sensoren auslösen zu können, wird der bisher genutzte Mule ESB durch das Business Rules Management System Drools 6.2 mit der RETE-Implementierung PHREAK ersetzt und die Funktionalität erweitert.

Schlüsselwörter

Complex Event Processing, Internet of Things, Complex Rules

CR-Kategorien

C.0 [**Computer Systems Organization**]: General—*System architectures*
; F.2.0 [**Analysis of Algorithms and Problem Complexity**]: General

*

Betreuer Hochschule: Prof. Dr.-Ing. habil. Natividad Martínez Madrid
Hochschule Reutlingen
Natividad.Martinez@Reutlingen-University.de
Betreuer Firma: Dipl. Wirt.-Ing. Eike Rehder
Fraunhofer IPA
Eike.Rehder@IPA.Fraunhofer.de

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

1 Einleitung

Sense&Act ist eine Internet-of-Things-Anwendung für den Einsatz im Industrie 4.0 Umfeld. Produktionsstätten sollen nachträglich mit Sensoren und Aktoren ausgestattet werden, um Vorteile der digitalen Vernetzung nutzen zu können. Sensoren werden an Maschinen angebracht, um bspw. Warenbewegungen oder Verschleißteile zu beobachten und wenn nötig eine Aktion auszulösen. Diese können Hinweise an Personal, wie die Erstellung eines Wartungstickets, oder Anweisungen an Maschinen, wie das automatische Befüllen eines Verbrauchgutes beinhalten. Sensoren sollen von Industrie-Partnern entwickelt oder in bestehende Systeme über Softwareerweiterungen bestehender Schnittstellen integriert werden. Ebenso können bestehende Web-Schnittstellen mithilfe eines Message Service Bus, wie das beim Fraunhofer IPA in Entwicklung befindliche Virtual Fort Knox System [9], angesprochen werden. Regeln werden von Nutzern in einem Web-Interface erstellt und verwaltet, dessen Fokus es ist, für nicht-Techniker einfach bedienbar zu sein. Im Zuge der wissenschaftlichen Vertiefung im Sommersemester 2015, sollte diesem System die Möglichkeit zur komplexen Regelausführung gegeben werden, um Daten verschiedener Sensoren in Kontext bringen zu können.

2 Anwendungsszenario

Das zu erweiternde System nutzt für die Regelausführung Mule ESB. Einen Java-basierten, weborientierten Enterprise Service Bus der über Port, Pfad und URL-Parameter Informationen an Anweisungslisten übergibt. Diese http-Endpoints initiieren, wenn die Bedingungen mit den Regeln übereinstimmen, eine vorher definierte Aktion. Mule besitzt kein Working Memory, kann aber über History-Werte aus der lokalen Datenbank Wertvergleiche anstellen, um Differenzen erkennen zu können. Das erklärte Ziel von Mule ESB ist, es Business-Prozesse zu automatisieren und Schnittstellen verschiedener webbasierter Applikationen zu verbinden. Ein oft genutztes Beispiel wäre eine Bestellung über ein Kundenportal, welches die Informationen an den Mule ESB sendet, der wiederum die Nachricht für die Verarbeitung in Salesforce oder SAP umwandelt und daran weiterleitet. Weiterhin wird ein Auftrag für den Versand generiert, so dass redundante Schritte für das Unternehmen entfallen. Die Wege verschiedener Datensätze und Aktionen sind somit fest vordefiniert.
Sense&Act nutzt dieses System um beim Eingang von Sensordaten zu entscheiden, ob diese eine Aktion auslösen sollen und wenn ja, diese Aktion zu initiieren. Jeder Sensor ist einer Ausführgruppe zugeteilt, die denselben Port belegen und an ihrem Pfad unterschieden werden. Somit besitzt jeder Sensor serverseitig eine Reihe an Regeln, die ausgeführt werden können.

3 Stand der Wissenschaft

Obwohl die Anfänge der modernen Complex Event Processing in den 70er Jahren liegen [6], ist es weiterhin ein aktives Thema in der Wissenschaft. Anwendung findet es unter anderem im Bereich Ereignisorientierte Simulation, Netzwerkmanagement und Active Databases. In den letzten Jahren findet man Complex Event Processing jedoch hauptsächlich im Business Management in Middlewareanwendungen zur Businessprozessautomatisierung [11]. So setzen Meyer et al. [12] ein Regel-System ein, um Service Level Agreements automatisiert in einem Flugkontrollszenario einzuhalten. Ein Problem das weiterhin besteht ist die Skalierung von Regel-Systemen. Zwar gibt es Ansätze wie von Rosenberg und Dustdar [14] oder von Kumarasinghe et al. [8], diese werden jedoch von der Industrie bisher noch nicht angenommen [2]. Wie in vielen anderen Gebieten wird versucht, eine standardisierte Sprache zu definieren. Das W3C Konsortium Rule Interchange Format (RIF) hat seinen Standard im Juli 2010 verabschiedet. Das RIF wurde konzipiert als Kommunikationssprache zwischen Regel-Systemen, ist jedoch auch eine vollwertige Regel-Sprache [7]. Ein konkurrierender Standard ist Production Rule Representation (PRR). Anstatt einen eigenen Standard zu etablieren, versucht die Event Processing Technical Society (EPTS) eine “Koordinierung und Harmonisierung” zu erreichen [4] unter anderem mithilfe eines Glossars für Complex Event Processing.

4 Komplexe Regeln

> ”Ein Ereignis repräsentiert etwas das stattfindet, passiert oder die aktuelle Sachlage verändert. [...] Einfache Ereignisse werden zusammengefasst, abhängig ihrer zeitlichen, kausalen und semantischen Beziehungen” [1]

Was passiert jedoch, wenn die Daten mehrerer Sensoren kombiniert werden sollen, um Ereignisse zu starten. Als Beispiel wird hier das Konzept einer verteilten Wetterstation genutzt. Windstärkesensor, Helligkeitssensor, Feuchtigkeitssensor, Barometer und Thermometer werden als physisch getrennte Geräte an jeweils geeigneten Orten aufgestellt. Stark ansteigender oder fallender Luftdruck dient als Indikator für einen bevorstehenden Wetterumschwung. Registriert der Lichtsensor zusätzlich einen starken Helligkeitsabfall und der Windsensor eine erhöhte Windgeschwindigkeit, kann ein Unwetter bevorstehen. Liefern mehrere Sensoren Indikatoren für schlechtes Wetter, sollen Fenster geschlossen, Jalousien hochgefahren und wenn nicht in Benutzung, offene Tore geschlossen werden.

4.1 Merkmale

Diese Beschreibung stellt wichtige Anforderungen an ein System zur Ausführung komplexer Regeln:

- Sie benötigen Fakten aus unterschiedlichen Quellen, die zu unterschiedlichen Zeitpunkten eintreffen.
- Es muss einen zeitlichen Kontext geben, um Relevanz von Fakten bewerten zu können.
- Es muss definiert sein, wie eine Regel ausgeführt wird.
- Es muss definiert sein, wie Fakten erstellt werden können.

Der Unterschied von Event Processing zu Complex Event Processing ist die Anzahl von Quellen pro auszuführender Regel [5]. Um Fakten aus diesen unterschiedlichen Quellen vergleichen zu können, muss ein Working Memory vorhanden sein, welches Regeln und Fakten über den aktuellen Zustand enthält.
Das System muss ein Zeitautomat (Echtzeitverhalten) sein, um Fakten kontinuierlich in Relation bringen zu können. So ist es nicht sinnvoll, die Helligkeitsdaten von vor einer Stunde mit den aktuellen Luftdruckwerten zu verknüpfen, da sich der Himmel seit den letzten Daten signifikant verdunkelt haben kann. Es muss somit einen zeitlichen Kontext geben der die Relevanz von Fakten bestimmt.
Weiterhin ist die Zeit nicht nur im Kontext relevant sondern auch in der Ausführung. Regeln müssen beinhalten, wie und wann sie ausgeführt werden. Da Konditionen über einen längeren Zeitraum übereinstimmen können, kann dies zu einer Vielzahl von Aktionen führen, obwohl nur eine ausgeführt hätte werden sollen.
Um beispielsweise Statusänderungen im System abzubilden wie, dass die Fenster geschlossen sind oder es Nacht ist, müssen Regeln in der Lage sein, Fakten zu erstellen. Diese können wiederum dazu führen, dass andere Regeln ausgeführt werden. Jedoch bietet dies die Möglichkeit für Endlosschleifen und Speicherüberläufe, was im System verhindert werden muss.

5 Ausführen kompl. Regeln

Das Ausführen komplexer Regeln kann auf unterschiedliche Weise geschehen. Die einfachste Variante wäre eine Datenbank die alle einkommenden Events und Regeln enthält und ein Stück Software, dass bei jedem Eintreffen neuer Events jede Regel ausführt und überprüft ob sie nun zutrifft. Dieser Ansatz ist nicht effizient, da auch Regeln geprüft werden deren Ausführung gar keinen Zusammenhang mit dem eintreffenden Event besitzen. Effizienter ist ein Ansatz, der durch Eigenschaften des eintreffenden Events ausschliesslich relevante Regeln überprüft und wahrscheinlicheren Regeln eine höhere Priorität zuweist. Da für das überprüfen komplexer Regeln die Status mehrere Sensoren notwendig sind, ist ein Speicher-Ansatz mit einer klassischen Datenbank, die auf Schreib- und Leseoperationen von Festplatten angewiesen ist, nicht zeit effizient, weshalb eine schnellere Speicherlösung für das Working Memory wie RAM verwendet werden sollte. Da Regeln von Nutzern auf Bedarf erstellt werden, müssen diese zur Laufzeit in das CEP-System eingetragen werden können. Ein Neustart des Systems soll hierbei umgangen werden, da dies Zeit benötigt, in der Events eintreffen und Regeln ausgeführt werden könnten.

5.1 RETE

Moderne Business Rule Engines verwenden den Rete-Algorithmus, der Fakten durch Musterabgleiche mit Regeln vergleicht. Regeln sind durch eine Left-Hand-Side und Right-Hand-Side definiert. Erst wenn alle Bedingungen auf der Left-Hand-Side erfüllt sind, wird die Right-Hand-Side ausgeführt.

Der Rete-Algorithmus wurde von Charles Forgy 1979 entwickelt und erstellt auf Basis der vorhandenen Regeln einen Entscheidungsbaum, der aus zwei Arten von Knoten besteht. α-Knoten sind Selektionsbedingungen, die sich auf einzelne Elemente im Working Memory beziehen und β-Knoten, die Bedingungen miteinander verknüpfen. [6] Die Regelausführung wird optimiert durch das Einschränken der zu überprüfenden Bedingungen. Identische Bedingungen werden verbunden und müssen somit nur einmal überprüft werden. Im Falle von Sense&Act wäre dies unter anderem der Sensor-Identifikator, der in jeder Regel überprüft wird. Der Rete-Algorithmus hat seit seiner Veröffentlichung mehreren Revisionen erfahren, so dass aktuelle Implementationen einen Leistungsgewinn von 500% vorweisen können. [13]

5.2 Drools / PHREAK

Drools ist ein Open Source Business Rules Management System und die frei verfügbare Variante des JBoss Enterprise BRMS. Seit Version 6 arbeitet Drools nach dem PHREAK-Algorithmus, welcher eine Weiterentwicklung des ReteOO-Algorithmus ist. Erweiterungen gegenüber aktuellen Rete-Varianten sind lazy-Verhalten, das ermöglicht mehr Regeln und Fakten im Working Memory zu verwalten, aber mehr Speicher benötigt, Agendagroups bei denen Regeln erst verglichen werden, wenn die Gruppe übereinstimmt und Drools-Spezifische Anpassungen wie Konfliktresolution durch Priorisierung und Regel-Attribute. Diese geben erweiterte Möglichkeiten für die Verhaltenssteuerung von Regeln, wie die Nutzung von Kalendern oder Zeitsteuerung. [10]

6 Konzeption

Für die Konzeption der Complex Event Processing Komponente mussten bestehende Voraussetzungen, die sich aus dem Sense&Act Anwendungsfall ergeben, erfüllt werden:
Die Kommunikation mit dem System findet über eine Web-Schnittstelle statt. Dies hat den Vorteil, dass Sensoren und Aktoren sich nicht in denselben Netzwerken befinden müssen und über bestehende Internet Infrastruktur kommunizieren können.
Das Working Memory muss mehrere hundert bis mehrere tausend Sensoren und Regeln beinhalten können, wenn ein Fabrik-Gebäude komplett mit Sensoren und Aktoren ausgestattet werden soll. Das Eintreffen von Sensordaten muss effizient und zeitnah erfolgen, da durch Push-Sensoren mehrere hundert Statusmeldungen pro Sekunde eintreffen können.
Regeln müssen zur Laufzeit hinzugefügt werden können, ohne dass das Regel-System neu starten muss oder wenn dies nicht möglich ist, neustarten ohne das Working Memory zu verlieren. Weiterhin dürfen keine eintreffenden Sensordaten verloren gehen, da sonst Gefahr besteht, selten eintreffende Sensorupdates zu verlieren und somit kritische Fakten nicht im Working Memory zu halten.
Da Sensoren unterschiedliche Anzahl und Typen von Informationen senden, muss das Verhalten von Datentypen, aber nicht die Anzahl im System fest definiert sein.
Das Sense&Act Konzept sieht vor, dass Sensoren von Industrie-Partnern entwickelt werden sollen, weshalb die Sensor-Kommunikation einfach umzusetzen und wenn möglich ein proprietärer Standard vermieden werden soll.

6.1 Konzeption

Im Vornherein fiel die Entscheidung für Drools als Regel Management System. Zwar nutzen die meisten Regel Management Systeme eine aktuelle Version des RETE-Algorithmus und bieten ähnliche Funktionalitäten, jedoch bietet Drools das in einer einfach zu integrierenden Bibliothek mit einer ausführlichen Dokumentation und einer Vielzahl an Helferfunktionen. Wie im vorherigen Kapitel dargelegt, arbeitet der verwendete PHREAK-Algorithmus schnell aber nicht speichereffizient, was im Sense&Act UseCase jedoch kein ernster Nachteil ist.

Für die Abbildung eines Events wurde das von IBM spezifizierte Common Base Event verwendet [3], da es der in Sense&Act bisher genutzten Sensor-Kommunikation sehr ähnelt und alle notwendigen Parameter-Typen unterstützt. Da jedoch davon ausgegangen wird, dass Sensoren ohne vorherige Anmeldung im System Events senden können, muss das Common Base Event soweit erweitert werden, dass der Parameter-Typ zwingend angegeben werden muss. So können Parameter wie Geo-Koordinaten oder Vektoren gesendet und von der Complex Event Processing Engine als solche erkannt werden. Dies verletzt, die in [3] spezifizierten Datentypen nicht, da Parameter weiterhin als String, Float oder Integer dargestellt werden.
Für die Sensor-Kommunikation wurde das offene und weitverbreitete JSON-Format ausgewählt. Sensoren müssen ein festgelegtes Objekt mit den Attributen Name, ID und einer Liste mit Parametern senden, wobei Parameter den jeweiligen Namen, Typ und Wert enthalten. Da bestehende Sensoren weiterhin verwendet werden sollen können, wird deren URL-Parameter basierte Sensor-Kommunikation ebenso unterstützt wie die JSON-Payload Kommunikation. Der Mule ESB wird zwar in der Regelausführung ersetzt, jedoch als Konnektor zu verschiedenen Aktoren wie Email oder SAP beibehalten.
Drools nutzt Fakten für sein Knowledge is Everything-System, für deren Nutzung erschlossen sich zwei Möglichkeiten: Ein Fakt pro Sensor, der bei neuem Eintreffen von Informationen ein Update erhält oder jedes Event ein Fakt, das über Zeit an Relevanz verlieren. Drools bietet hierfür einen Sliding-Window Modus, der jedoch Probleme aufwarf, weshalb für die Variante ”ein Fakt pro Sensor” entschieden wurde.
Um Fakten im Falle eines Hardware- oder Softwareversagens wiederherstellen zu können, soll das Working Memory in regelmäßigen Abständen in einer Datenbank gesichert und beim Neustart automatisch wiederhergestellt werden. Um für denselben Fall einkommende Sensor-Kommunikation nicht zu verlieren, soll eine Message Queue auf einer separaten Maschine eingerichtet werden. Im Falle eines Ausfalls kann diese, für einen durch die Cache-Größe definierten Zeitraum, Nachrichten zwischenspeichern.

7 Umsetzung

Das CEP-System wurde auf Basis von Drools 6.2 umgesetzt. Hier wurde eine Stateful Session als Working Memory gewählt und zu Beginn die Sliding-Window Methode gewählt, die jedes Sensor-Update als separaten Fakt im Working Memory behandelt. Eigentlich sollten, nachdem eine definierten Anzahl an Fakten desselben Sensors im Working Memory vorhanden sind, die veralteten Fakten gelöscht werden. Da dies jedoch nicht der Fall war, wurde auf die Sliding-Window Methode verzichtet.
Für Kommunikation mit dem Modul wurde eine Web-Schnittstelle auf Basis von Jetty, Jersey, und RSX implementiert. Ein http-Endpoint für die Sensoren und eine RESTful API für das Erstellen von Regeln und Abfragen von Statusinformationen. Da vorerst nur die bestehenden Sensoren mit dem System arbeiten, wurde der Fokus auf die URL-Parameter Schnittstelle gelegt. Der JSON-Payload-Endpoint wurde für zukünftige Verwendung ebenso implementiert, der auf POST-Requests reagiert. Aus beiden Endpoints werden Common Base Events generiert und als Fakten in das Drools Working Memory gelegt. Sendet ein Sensor zum ersten Mal seinen Status an das System, wird ein neuer Fakt im Working Memory angelegt. Jegliche Statusupdates danach, initiieren einen Update des bestehenden Fakts und einen Abgleich mit den vorhandenen Regeln.
Für das Anlegen und Ändern von Regeln wurde nach mehreren Versuchen, diese zur Laufzeit zum Working Memory hinzuzufügen oder daraus zu löschen, die Methode gewählt, eine gepackte Java-Ressource nachzuladen. Diese hat an sich die Funktion, eine aktualisierte Version aus einem Repository zu laden, kompilieren und zur Laufzeit nachzuladen. Dies ermöglicht ein automatisiertes, zentralisiertes Versionsmanagement über mehrere Instanzen und Con-

tinuous Delivery. Da das Projekt mit dem Jetty-Webserver, Maven und mehreren zusätzlichen Bibliotheken zwischen 400 und 800 Megabyte Arbeitsspeicher benötigt und Regeln relativ häufig geändert werden sollen, wurde ein separates RulesLoader Projekt erstellt, dessen Ressourcen vom CEP-System erstellt werden und das nur einen Bruchteil der Größe besitzt. Hierfür werden auf Basis eines Regel-Templates eine Drools-Regel erstellt und die in Attributes, Left-Hand-Side und Right-Hand-Side getrennt in einer lokalen Datenbank gespeichert. Daraufhin wird aus allen in der Datenbank befindlichen Regeln eine Drools Rule Language Datei generiert und im RulesLoader Projekt in den Ressourcen abgelegt. Dieses Projekt wird dann von Maven erstellt und im lokalen Maven-Repository hinterlegt und von dort vom CEP-System als neue Version nachgeladen. Da das Rules-Loader Projekt nur aus den für das Kompilieren relevanten Klassen und der DRL-Ressource besteht, wird die aktuelle Regel-Ressource überschrieben und ausschließlich die Regeln im Working Memory überschrieben.
Die Fakten Sicherung wird in dieselbe Datenbank geschrieben, aus der auch die DRL-Ressource generiert wird und zum Systemstart als Common Base Events wieder in das Working Memory geladen. Auf die MessageQueue wurde vorerst, aus Mangel an separaten Servern, verzichtet.

7.1 Probleme

Das wiederkehrende Problem bei der Umsetzung der CEP-Komponente, war die teils veraltete Dokumentation von Drools 6.2 und den Wechsel von Version von 5 auf 6 im vergangenen Jahr. So wurde das Deployment Modell komplett auf Maven umgestellt und die notwenigen Änderungen aber nur teilweise in der Dokumentation behandelt. Weiterhin wurde das Session-Konzept vereinheitlicht, durch das Umbenennen und Umorganisieren von Klassen und Methoden, was jedoch in der Dokumentation noch nicht vollständig geändert wurde. So werden Verweise auf Architektur-Bereiche aus Version 5 erwähnt, die in Version 6 ohne Modifikationen nicht möglich sind. Zudem wurden Erweiterungsmodule noch nicht auf die aktuelle Version portiert, aber ebenso nicht aus der Dokumentation entfernt, wie das Quartz-Kalender Modul, das das Verwenden von Wochentagen oder Uhrzeiten als Regel-Attribute ermöglichen. Dies konnte jedoch durch Foreneinträge und dem Erstellen einer Adapterklasse gelöst werden.

8 Fazit und Aussicht

Es wurde eine Complex Event Processing Komponente für das IoT Industrie 4.0 System Sense&Act auf Basis von Drools 6.2 konzipiert und implementiert. Durch die effizientere Abarbeitung der eintreffenden Events als mit der bisherigen Event Processing Komponente und der Erweiterung durch komplexe Regeln, ist das System performanter und vielseitiger. Durch die Nutzung des CommonBaseEvents und Übertragung per JSON-Payload können detailliertere Informationen von Sensoren für die Regelausführung verwendet werden. Mit der Verwendung einer Message Queue und Drools 6.2 wird das Ändern der Regelbasis zur Laufzeit ermöglicht, ohne dass ein Neustart des Systems notwendig ist und Events verloren gehen.
Da Sense&Act aktuell ausschließlich für Demonstrationszwecke mit nur einer geringen Anzahl von Sensoren und Aktoren verwendet wird und das Frontend das Erstellen von Komplexen Regeln nicht unterstützt, ist ein Umstieg auf die CEP-Komponente aktuell nicht notwendig. Jedoch ist eine komplette Überholung des Frontends und der Architektur und Komponenten auf Basis von Microservice Architektur für Platform-as-a-Service-Deployment in Docker-Container in Arbeit, in dessen Zentrum die vorgestellte Complex Event Processing Komponente arbeiten wird.

Literatur

[1] D. Anicic, S. Rudolph, P. Fodor, and N. Stojanovic. Stream reasoning and complex event processing in ETALIS. *Semantic Web*, 3(4):397–407, 2012.

[2] A. Buchmann and B. Koldehofe, editors. *IT-Information Technology*, volume 51. Oldenbourg Verlag, October 2009. Online verfügbar unter http://www2.informatik.uni-stuttgart.de/cgi-bin/NCSTRL/NCSTRL_view.pl?id=BOOK-2009-02 Besucht am 10.11.2015.

[3] I. Corporation. Canonical situation data format: The common base event v1. 0.1, Apr. 2003. Online verfügbar unter http://eclipse.org/tptp/platform/documents/resources/cbe101spec/CommonBaseEvent _SituationData_V1.0.1.pdf ; Besucht am 26.10.2015.

[4] M. Eckert and F. Bry. Aktuelles schlagwort "complex event processing (cep)". Online verfügbar unter https://epub.ub.uni-muenchen.de/14902/1/bry_14902.pdf Besucht am 10.11.2015, 2009.

[5] M. Eckert and F. Bry. Complex event processing (cep). *Informatik-Spektrum*, 32(2):163–167, 2009. Online verfügbar unter http://dx.doi.org/10.1007/s00287-009-0329-6; Besucht am 26.10.2015.

[6] C. L. Forgy. Expert systems. chapter Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem, pages 324–341. IEEE Computer Society Press, Los Alamitos, CA, USA, 1990. Online verfügbar unter http://dl.acm.org/citation.cfm?id=115710.115736; Besucht am 26.10.2015.

[7] W. R. W. Group. Rif wiki. Website. Online verfügbar unter http://www.w3.org/2005/rules/wiki/RIF_FAQ; Besucht am 11.11.2015.

[8] T. Holvoet and M. Viroli. *Coordination Models and Languages - 17th IFIP WG 6.1 International Conference, COORDINATION 2015, Held as Part of the 10th International Federated Conference on Distributed Computing Techniques, DisCoTec 2015, Grenoble, France, June 2-4, 2015, Proceedings*. Springer, Berlin, Heidelberg, 2015.

[9] F. IPA. Virtual fort knox. Website. Online verfügbar unter https://www.virtualfortknox.de; Besucht am 26.10.2015.

[10] JBoss-Drools-team. Drools documentation - version 6.2.0.final, Mar. 2015. Online verfügbar unter https://docs.jboss.org/drools/release/6.2.0.Final/drools-docs/pdf/drools-docs.pdf; Besucht am 26.10.2015.

[11] D. C. Luckham. A short history of complex event processing. part 1: Beginnings. 2007. Online only http://complexevents. com/wp-content/uploads/2008/02/1-a-short-history-of-cep-part-1.pdf.

[12] F. Meyer, R. Kroeger, and M. Milekovic. An approach for knowledge-based IT management of air traffic control systems. In *2013 9th International Conference on Network and Service Management (CNSM)*, pages 345–349, October 2013.

[13] J. Owen. Worlds fastest rules engine. Website, sep 2010. Online verfügbar unter http://www.infoworld.com/article/2626208/application-development/world-s-fastest-rules-engine.html; Besucht am 26.10.2015.

[14] F. Rosenberg and S. Dustdar. Towards a distributed service-oriented business rules system. In *Web Services, 2005. ECOWS 2005. Third IEEE European Conference on*, Nov 2005.

Skalierbarkeit von Online-Lernsystemen

Raphael Fritsch
Reutlingen University
Raphael.Fritsch@Student.Reutlingen-University.DE

Abstract
In dieser Ausarbeitung wird im ersten Teil aufgezeigt, was Online Lernsysteme sind und für was sie verwendet werden. Dann wird das verwendete Programm „Accelerator“ und die darunter liegende Multipoint Control Unit (MCU) in Aufbau und Funktionsweise beschreiben. Anschließend wird gezeigt, warum Peer to Peer Verbindungen für solche Systeme ungeeignet sind und wie die Anwendung, an bestimmten Stellen skaliert werden muss, um viele Clients zu unterstützen. Zum Schluss werden noch weitere Möglichkeiten aufgezeigt, wie eine Skalierung bei noch größeren Belastungen theoretisch implementiert werden müsste.

Schlüsselwörter
Online-Lernsystem, WebRTC, Lastverteilung, Skalierung

CR-Kategorien
Performance, Algorithms

1 Einleitung
Der Unterricht an den meisten Schulen und Universitäten setz immer noch Anwesenheitsplicht der Schüler und Studenten voraus. Mittlerweile wird jedoch neben herkömmlichen Methoden auch das Lehren über Videos, beziehungsweise interaktiv über Online Vorlesungen eingesetzt. Neben Sprache werden dort oft auch Präsentationen, Kamerabilder und Bildschirme übertragen. Da dieser Unterricht oft für mehr als 60 Personen gehalten wird, müssen entsprechende Interaktive System gut skalieren und zudem noch möglichst Plattformübergreifend funktionieren. In dieser Ausarbeitung soll gezeigt werden wie ein skalierbares System funktionieren kann und an welchen Stellen eine Skalierung überhaupt nötig ist.

2 Stand der Wissenschaft
Im Moment gibt es im Bereich Online Vorlesungen kaum Software die sich für die Lehre eignet und dabei Funktionen wie Präsentationen und Interaktionen mit Studenten vereint. Zu den Bekannteren Anbietern gehören u.a. Netviewer, Vitero oder Adobe Acrobat Connect Pro.

Diese haben jedoch meist sehr Kostenintensive Lizenzmodelle. Zudem muss bei den meisten Anbietern neben der Serversoftware auf jedem Client extra ein Programm installiert werden um an einer Konferenz teilzunehmen. Außerdem ist es oft schwer oder auch nicht Möglich diese Programme an eigene Bedürfnisse oder Wünsche anzupassen.

Deshalb wurde auf einer zur Verfügung stehenden OpenScource MCU ein eigenes Programm entwickelt.

Betreuer Hochschule: Prof. Dr.-Ing. Peter Hertkorn
Hochschule Reutlingen
Peter.Hertkorn@Reutlingen-University.de

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

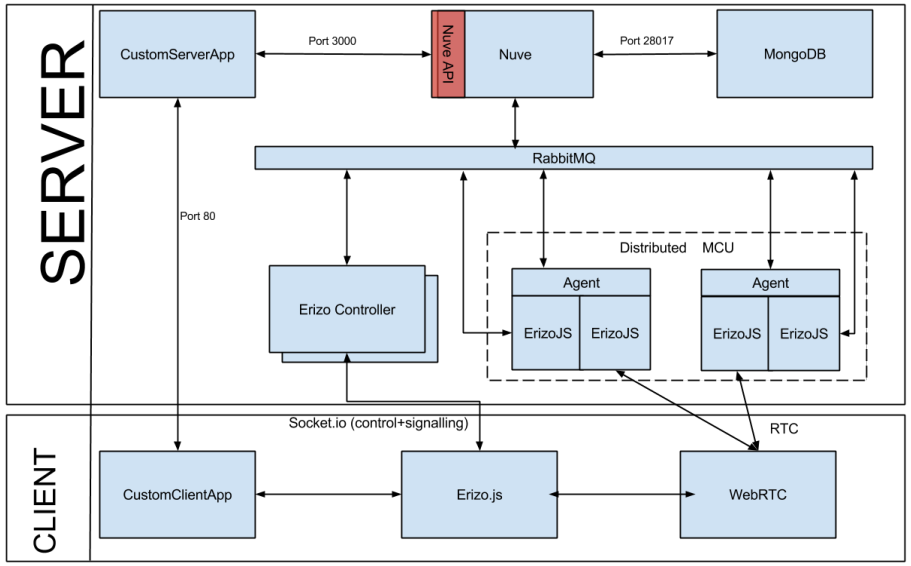

Abbildung 1: Licode Architektur

Diese MCU kann bei korrekter Konfiguration die Last auf mehrere Rechner verteilen ohne die Verwendung von bekannten Cloud Tools wie Docker oder OpenStack.

3 Accelerator

Der Accelerator ist eine Webapplikation die es ermöglicht Vorlesungen in einer Browserbasierten Umgebung zu halten, ohne zusätzliche Software auf Seiten der Benutzer installieren zu müssen. Diese Software wurde im Rahmen einer Vorlesung an der Hochschule Reutlingen entwickelt und wird seither auch dort eingesetzt. Durch eine Browserbasierte Lösung wird einerseits Plattformunabhängigkeit erreicht, aber auch leichte Erweiterbarkeit durch HTML5 Webtechnologie gewährleistet.

3.1 Architektur

Der Accelerator ist aufgeteilt in Client und Serverseite. Der Client besteht aus einer HTML/JavaScript Anwendung die sowohl über einen Websocket mit dem Accelerator Server kommuniziert, sowie Sprach- und

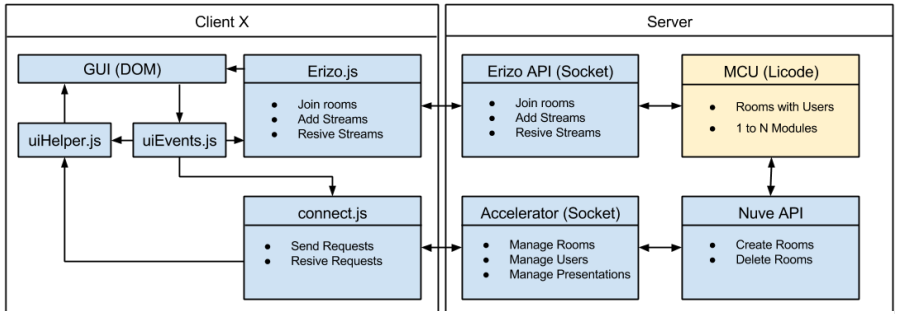

Abbildung 2: Accelerator Architektur

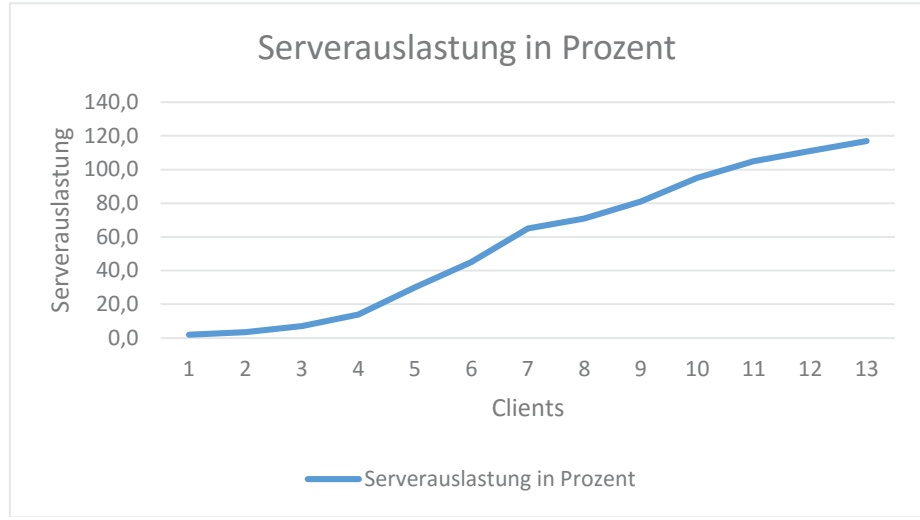

Abbildung 3: Server Auslastung auf einer CPU

Videodaten über WebRTC mit Licode austauscht.

Der Accelerator Server ist eine NodeJs [1] Anwendung die auf Licode [2] aufbaut. Licode ist wiederum eine Opensource Multipoint Control Unit (MCU) welche die Sprach- und Videostreams der Clients verwaltet und dafür sorgt dass jeder Client die richtigen Sprach- und Videodaten erhält.

In Abbildung 1 ist die Architektur von Licode dargestellt. Die Module „CustomServerApp“ und „CustomClientApp“ wurde dabei durch den Accelerator Server und Client ersetzt. Die sonstigen Module von Licode übernehmen folgende Aufgaben:

Das Nuve Modul: Stellt eine Schnittstelle zur Verfügung um per CustomServerApp neue Räume zu erstellen/löschen, Clients zu verwalten und Räume auf die zur Verfügung stehenden Erizo Controller zu verteilen.

MongoDB: Diese Datenbank wird verwendet um Raum Informationen und Tokens zu speichern. Hier werden jedoch keine Benutzerdaten verwaltet. Dies muss in der CustomServerApp implementiert werden.

Erizo Controller: Dieses Modul verwaltet Räume und startet ggf. neue Erizo Agents. Dieses Modul wird vom Nuve Modul automatisch erkannt wenn beide mit der gleichen RabbitMQ Instanz verbunden sind.

Erizo Agent: Dieses Modul ist die eigentliche MCU und Verteilt Sprach- und Videostreams. Es kann, genau wie der Erizo Controller mehrfach und auf unterschiedlichen Maschinen ausgeführt werden. Auch dieses Modul wird von Nuve erkannt, wenn beide mit der gleichen RabbitMQ Instanz verbunden sind.

Accelerator: Der Accelerator Server ist wie auf der rechts in Abb. 2 zu sehen aufgebaut. Er verwaltet Verbindungen von Clients und steuert die Nuve API an. Außerdem sorgt er dafür dass Präsentationen und Interaktionen der Benutzer korrekt koordiniert werden. Zudem bietet er einen Websocketserver an, über den die Clients Chatnachrichten austauschen können und sich über bestimmte Tasten beispielsweise melden oder klatschen können.

Der Accelerator Client ist die Weboberfläche (Abbildung 2 links) die dem Benutzer vom Server zur Verfügung gestellt wurde. Der

Client baut sowohl eine Verbindung zum Accelerator Server auf, wie auch WebRTC Verbindungen zur MCU. Des Weiteren bietet er das Interface für Präsentationen, Bildschirmübertragung, Textnachrichten, Zeichenfelder und andere Interaktionsmöglichkeiten.

4 Technische Umsetzung

Der Accelerator ist eine Webanwendung und wurde mit HTML und Javascript umgesetzt. Für die Verbindung zum Server wird sowohl eine Websocketverbindung für Interaktionen, als auch eine WebRTC Verbindung für Medienstreams aufgebaut.

4.1 WebRTC

WebRTC (Web Real-Time Communication) ist ein offenes Projekt welches Browsern und Mobilen Anwendungen Echtzeitkommunikation über eine einfache API ermöglicht. [vgl. 5] WebRTC ermöglicht die Kommunikation ohne zusätzliche Plugins muss aber im Browser selbst implementiert sein. Dies ist für die meisten Browser mittlerweile gegeben. So wird WebRTC von Browsern wie *Google Chrome*, *Mozilla Firefox*, *Opera* und vielen weiteren Unterstütz. Keine Unterstützung besteht derzeit (10/2015) noch für die großen Browser *Internet Explorer* und *Safari*. Eingeschränkte Unterstützung wird vom neuen Microsoft Browser *Edge* angeboten. [6] WebRTC ist eine Weiterentwicklung von WebSockets die zwar eine Pluginfreie Kommunikation ermöglichen, jedoch ohne Echtzeitanspruch funktionieren. [7, 21]

WebRTC ist in der Lage mit Hilfe von Interactive Connectivity Establishment (ICE) Verbindungen durch Network Address Translations (NAT) und Firewalls hindurch aufzubauen und aufrecht zu erhalten. Dazu wird ein extra ICE Server verwendet. [7,25] Dies gilt sowohl für die Verbindung vom Client zum Server als auch eine direkte Verbindung von Client zu Client (Peer to Peer) Im Internet gibt es viele kostenlose ICE Server die für diesen Zweck verwendet werden können. Obwohl über diesen Server nur eine Verbindung aufgebaut wird und nicht die eigentlichen Daten versendet werden, sollte bei sensiblen Gesprächen bzw. Daten ein eigener ICE Server installieren werden.

4.2 Serverseitige Umsetzung

Der Server ist, wie oben beschrieben, eine NodeJS Anwendung die auf Javascript basiert.

Der Accelerator Server verwendet das NodeJS Module „Express“ und stellt darüber einen Webserver zur Verfügung welcher den Accelerator Client zur Verfügung stellt. Die Verbindung wird über eine Verschlüsselte Leitung (https) aufgebaut. Zum einen, weil Screensharing nur in einer Verschlüsselten Umgebung funktioniert. Zum anderen, weil Verschlüsselung keinen Mehraufwand für die Benutzer bedeutet. Zur Verwaltung von ZIP komprimierten Daten wurde zudem das Modul „yauzl“ [8] verwendet. Dadurch ist es möglich gepackte Daten, wie Präsentationen, als eine Datei hochzuladen und zu verarbeiten.

Auch ein großer Teil von Licode sind NodeJS Module welche dann über eine API sowohl vom Accelerator Client als auch vom Server aus angesteuert werden können. Der Kern von Licode ist jedoch die MCU (Erizo Agent) welche in C++ programmiert wurde.

4.3 Peer to Peer vs. MCU

Die kürzeste Verbindung zweier Punkte ist die gerade. So verhält es sich nicht nur bei Punkten im zweidimensionalen Raum sondern auch bei zwei Rechnern die über das Internet kommunizieren wollen. Bei einer direkten Verbindung ist sowohl das Fehlerpotenzial als auch die Antwortzeit (ping) geringer. Bei zwei Clients bietet es sich deshalb an, eine direkte Verbindung (Peer to Peer) aufzubauen und darüber zu kommunizieren. Werden es jedoch mehr als zwei Clients, bildet sich ein so genanntes „Netz“ (Mesh). Das bedeutet wenn jeder der drei mit den jeweils zwei anderen im „Netz“ kommunizieren will, muss die doppelte Datenmenge verschickt werden (an jeden anderen Client dasselbe Packet). Diese Methode ist jedoch sehr ungünstig, vor allem bei sehr vielen Clients oder einer entsprechend langsamen Upload Geschwindigkeit. Bei einem Videochat ist es deshalb schwer möglich mit Uploadgeschwindigkeiten von unter 1 MB/s je nach Video Qualität mehr als zehn andere Clients gleichzeitig mit Videodaten zu versorgen.

An diesem Punkt kommt die Multipoint Control Unit (MCU) ins Spiel, welche die Aufgabe des Verteilers in der sogenannte Stern Topologie erfüllt. Dabei verbinden sich alle Clients nur mit diesem einen Server anstatt viele Verbindungen untereinander aufzubauen. Der Server bietet dafür „Räume“ an, die die jeweiligen Clients betreten können um dort ihre Daten auszutauschen. Das bietet den Vorteil, dass jeder Client seine Daten nur einmal hochladen muss, egal wie viele andere Clients er versorgen will und verschiebt die Rechenlast vom Client auf den Server, was für schwache Mobile Geräte von Vorteil ist.

5 Lastverteilung

Jeder Client erzeugt eine neue Last für den MCU Server. Der Server muss dabei eingehende Video- und Audiostreams entgegennehmen, kopieren und an alle angemeldeten Clients im gleichen Raum verteilen. Wie in Abbildung 3 zu sehen ist, steigt die Serverauslastung bei Verwendung einer 2,5 GHz CPU Schon bei mehr als 10 Benutzern über 100%. Das System soll jedoch Spitzen von mehr als 60 Clients verarbeiten können. Deshalb muss die MCU skaliert werden. Dabei ist zu beachten, dass die Auslastung bei größerer Anzahl an Clients nichtmehr Liniar ansteigt, da jeder MCU Prozess bei neuen Verbindungen die Steamdaten an sehr viele andere Clients schicken muss.

Licode kann über die Änderung der Konfiguration auf mehr als eine CPU skaliert werden. Damit kann die Last theoretisch beliebig weit verteilt werden. So können auf einem Server mit 16 CPUs schon mindestens 60 Clients verarbeitet werden. Je mehr Clients sich jedoch verbinden desto schneller steigt die Auslastung und die Wahrscheinlichkeit einer Überlastung des Servers. Um mehr Clients verarbeiten zu können muss die MCU auf mehrere Server aufgeteilt werden. Deshalb wurde im Laufe dieser Arbeit der AccController erstellt, welcher es ermöglicht die MCU einfach auf mehrere Server zu verteilen.

6 Erläuterung des Installierten Systems

Um das System einfach skalierbar zu machen, und zusätzlich die Verfügbarkeit der Anwendung möglichst hoch zu halten, wurde der AccController im Zuge dieser Arbeit erstellt.

Der AccController ist wiederum in Hauptserver (AccController) und Client (DistAccClient) aufgeteilt. Der Server stellt zur Verwaltung des Accelerator Haupt- und der verteilten Server ein Webinterface zur Verfügung. Darüber können alle Services überwacht werden. Dabei werden sowohl die letzten Status- und Fehlermeldungen angezeigt, aber auch die letzte Startzeit. Zusätzlich kann jeder Service einzeln gestartet und gestoppt werden. Zu jedem Server wird die Auslastung in Prozent angezeigt. So kann entschieden werden, ob ein neuer Server dazu geschaltet werden muss. Wichtig ist dabei, dass neu gestartete verteilte Server nicht direkt die Last der laufenden Server übernehmen können. Sie übernehmen nur die Last von neu verbundenen Clients.

Welche Services von jedem Server zur Verfügung stehen wird in der „services.conf“ festgelegt. Dort kann neben dem Pfad zum gewünschten Service auch festgelegt werden, ob dieser Service auch von verteilten Servern angeboten werden soll. Der Accelerator Server muss z.B. nicht auf verteilten Servern zur Verfügung stehen. Die services.conf ist eine JSON Datei welche die verschiedenen Services als Objekte eines Arrays beinhaltet.

Jeder Service muss eine NodeJS Anwendung sein.

Beispiel eines Services:

Quellcode 1: Konfigurationsbeispiel eines Services

```
{
  "desc" : "Licode - ErizoAgent",
  "name": "ea",
  "dir" : "../erizo_controller/erizoAgent.js",
  "distributed" : true,
  "options": {
   "cwd":"../../erizo_controller/erizoAgent/",
   "max": 1,
   "silent": true,
   "args": []
  }
}
```

Desc: Beschreibt den Service auf der Oberfläche des AccControllers. In diesem Fall handelt es sich um den Erizo Agent.

Name: Dabei handelt es sich um den internen Namen des Service. Dieser wird nur im Programmcode verwendet und darf keine Leerzeichen enthalten.

Dir: Das Verzeichnis an dem sich der Service (NodeJS Anwendung) befindet.

Distributed: Legt fest ob dieser Service auch auf einem Verteilten Server zur Verfügung stehen soll. Dies ist deshalb wichtig, da die services.conf auch auf dem Hauptserver verwendet wird. Da der Erizo Agent auf verteilten Server verwendet werden soll steht dieser Wert auf „true“.

Options: Diese Optionen werden von forever-controller benötigt. Diese sind nicht im Code festgelegt damit die Werte für jeden Service individuell angepasst werden können.

Cwd: Zeigt auf den Oberordner der Anwendung, die gestartet werden soll.

Max: Die Maximale Anzahl an Iterationen, die ein Service durchlaufen soll.

Silent: Legt fest ob forever-monitor debug Meldungen für diesen Service ausgeben soll.

Args: Weitere Argumente die dem Service mit übergeben werden sollen.

Neben diesen Optionen können noch viele weitere Argumente übergeben werden (siehe https://github.com/foreverjs/forever-monitor)

Skaliert werden im Falle des AccDistClients nur Erizo Controller und Erizo Agent, welche dann über RabbitMQ mit Anfragen versorgt werden. RabbitMQ ist ein Messaging Service, welcher im nächsten Kapitel näher erläutert wird.

6.1 Messaging Queues

Messaging Queues ermöglicht es, anderen Programmen asynchron zu kommunizieren und dadurch auch ohne direkte Kommunikation miteinander zu interagieren. Dadurch ist es auch möglich, dass Anwendungen auf mehreren Servern verteilt werden können und über die Verbindung auf eine Messaging Queue Last vom Hauptserver genommen werden kann.

Mit Licode wird bereits RabbitMQ [3] mit installiert. Darüber kann die Last von mehreren Clients auf andere Server verteilt werden. Dies geschieht sobald sich ein neuer Erizo Controller Prozess oder Erizo Client Prozess auf RabbitMQ verbinden. Ab diesem Zeitpunkt werden neue Clients per scheduling (round robin) auf alle Verfügbaren Server verteilt. Bei der Verteilung des Erizo Controllers auf mehrere Server wird die Raumverwaltung auf mehrere Server verteilt. Das heißt sobald mehr als ein Raum erstellt wurde, wird die Verwaltung des zweiten Raums vom verteilten Server übernommen.

Für den Fall das der Erizo Agent auf einen anderen Server zusätzlich gestartet wird, wird der Kopierprozess von Sprach- und Bilddaten von diesem Server mit übernommen.

Da die Verwaltung von Räumen so gut wie keine Rechenleistung kostet, reicht es in der Regel den Erizo Agent auf mehrere Server zu verteilen. Nur aus Verfügbarkeitsgründen bietet es sich an auch den Erizo Controller zu verteilen, um so Ausfallsicherheit zu gewährleisten.

6.2 Umsetzung der Skalierung und Evaluation

Um den Hauptserver, auf dem Accelerator bereits installiert ist, zu skalieren muss Licode auf mindestens einem anderen Server installiert werden. Nachdem der AccControler auf den neuen Server kopiert wurde, muss die Ziel IP-Adresse so angepasst werden, damit sie auf den Hauptserver zeigt. Auf dem Hauptserver muss die Hauptkomponente des AccControllers gestartet sein unter den sich auch RabbitMQ befindet.

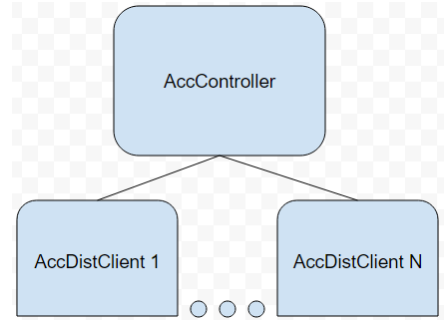

Abbildung 4: Strucktur des AccControllers mit Clients

Alle AccDistClients verbinden sich per TCP Verbindung mit dem AccController in einer Stern Topologie (siehe Abbildung 4). Über diese Verbindung können Services gestartet werden und Statusnachrichten von diesen übermittelt werden. Sobald über diese Verbindung der Erizo Agent Prozess gestartet wurde, verbindet er sich bei korrekter Konfiguration mit dem RabbitMQ Server. Wichtig zu beachten ist hierbei, dass auch verteilten Clients so eingestellt werden damit sie auf allen zur Verfügung stehenden Prozessorkernen rechnen. So steigt die Gesamtauslastung bei 20 Personen und verteilter Anwendung auf zwei Servern mit jeweils 16 CPUs auf nur 2,5% je Server. Da es in diesem Fall mehr CPUs als Personen gibt, kann jeder CPU eine Person zugeteilt werden.

6.3 Robustheit der Anwendung

Der Begriff Robustheit bezeichnet die Fähigkeit eines Systems, Veränderungen ohne Anpassung seiner anfänglich stabilen Struktur standzuhalten. [4] Dies heißt für die Anwendung des Accelerators, die Möglichkeit auf Fehler zu reagieren. Das bedeutet vor allem, dass die Serversoftware stabil läuft und bei Fehlern, die von einem Client aus hervorgerufen werden nicht direkt komplett den Service einstellt. Dies wird durch viele Try Catch Anweisungen im Programmcode gewährleistet, die dies verhindern können. Zusätzlich wird durch die Verwendung von forever gewährleistet, dass bei einem Absturz der Serversoftware der entsprechende Fehler protokolliert wird und der Service erneut gestartet wird. Dies gilt sowohl für den Hauptserver wie auch alle Services (Nuve, Erizo Controller, Erizo Client) die über den AccController gestartet wurden. Zusätzlich gilt dies auch für verteilt eingebundene Server.

7 Fazit und Ausblick

In dieser Arbeit wurde gezeigt wie das Online Lernsystem Accelerator und die darunterliegende Basis Accelerator skaliert werden kann. Dies wurde mit der Anwendung AccController realisiert, mit der verteilte Server verwaltet werden können und die Robustheit des Systems erhöht wird.

Als Verbesserung könnte neben den Manuellen Starts von verteilten Servern auch noch eine automatische Methode implementiert werden. Dies ist jedoch nur nötig wenn das System von vielen Gruppen und Benutzern benutz wird, oder sehr unausgeglichen belastet wird. Da sich AccDistClients automatisch beim Start auf

den AccController verbinden, muss theoretisch bei erhöhter Auslastung nur ein weiterer virtueller Server gestartet werden um die Last weiter zu verteilen.

Eine weitere Möglichkeit wäre die Integration von Cloud-Computing Software wie OpenStack in Kombination mit Docker. Die Integration in Docker wird seit 10. September von Licode selbst unterstützt und würde sich demnach anbieten.

Bei einer Belastung von unter 100 Benutzern ist die aktueller Konfiguration mit zwei Servern jedoch völlig ausreichen.

8 Literaturverzeichnis

[1] Node.js Foundation. Node.js. https://nodejs.org/en/. 2015. Accessed 22 October 2015.

[2] Pedro Rodriguez, Alvardo Alonso, Joaquin Salvachua. Lynckia: Web Realtime Communications Solutions. http://lynckia.com/. 2015. Accessed 22 October 2015.

[3] Pivotal Software, Inc. RabbitMq. https://www.rabbitmq.com. 2015. Accessed 22 October 2015.

[4] Wieland, A., Wallenburg, C.M. The ability of a [system] to resist change without adapting its initial stable configuration. In International Journal of Physical Distribution & Logistics Management, S. 42(10). 2012

[5] Google Inc. WebRTC. http://www.webrtc.org/. 2015. Accessed 22 October 2015.

[6] Adam Brault. &yet. http://iswebrtcreadyyet.com/. 2015 Accessed 22 October 2015.

[7] Altanai. WebRTC Integrator's Guide. Packt, 2014. ISBN 978-1-78398-126-7

[8] Thejoshwolfe. Yauzl. https://www.npmjs.com/package/yauzl, 2015. Accessed 22 October 2015.

Echtzeitsimulationen von Industriemaschinen und deren Verwendung zur virtuellen Inbetriebnahme *

Thomas Gulde
Reutlingen University
Thomas.Gulde@Student.Reutlingen-University.DE

Abstract

Im Zuge dieser Arbeit wird die Möglichkeit der Echtzeitsimulation von Industriemaschinen im Einsatzbereich der virtuellen Inbetriebnahme untersucht.
Diese Evaluierung soll die Verwendbarkeit der Software *Virtuos* zur Generierung eines solchen virtuellen Modells untersuchen.
Es gilt zu überprüfen, ob die Verwendung eines solchen Modells einen positiven Einfluss auf wichtige Faktoren wie Inbetriebnahmezeit, Softwarequalität und Testbarkeit von Industriemaschinen hat und somit den Aufwand rechtfertigt.
Diese Arbeit soll neben den verschiedenen Möglichkeiten des verwendeten Softwarepaketes die Integration solcher Modelle in den gesamten Entwicklungsprozess einer Anlage darstellen und auf diese Weise die Möglichkeiten und den Nutzen einer virtuellen Maschine offenlegen.

General Terms

Verification, Reliability, Security

*

Betreuer Hochschule: Prof. Dr.-Ing. Marcus Schöller
Hochschule Reutlingen
marcus.schoeller@Reutlingen-University.de
Betreuer Firma: Wolfram Schäfer
IT-Engineering GmbH
schaefer@ite-web.de
Informatics Inside 2015
Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

CR-Kategorien

I.6.5 [**Computing Methodologies**]: Simulation and modeling, Model Development

1 Motivation

Die Inbetriebnahme von neu entwickelten Maschinen ist ein wichtiger und elementarer Bestandteil des Entwicklungsprozesses. Während der Inbetriebnahme werden sämtliche Komponenten, wie Aktoren, Sensoren und sämtliche weiteren beteiligten Untersysteme einer Industrieanlage in den operativen Zustand überführt und für die Ansteuerung durch die Steuerungskomponenten, meist **S**peicher
Programmierbare **S**teuerungen (SPS), vorbereitet.
Ist diese Erstinbetriebnahme erfolgt, kann mit der eigentlichen Programmierung der Anlage, dem Hinterlegen von Regeln und Abläufen begonnen werden (vgl. [3], S. 1),
Diese Abfolge zu durchbrechen, gestaltet sich innerhalb des herkömmlichen Entwicklungsprozesses als unmöglich. Die meisten Programmierarbeiten lassen sich ohne vorhandene Hardware nicht erledigen und vor allem ist eine Testbarkeit der entwickelten Routinen nicht gegeben.
Neben dieser Tatsache zeigt die Realität, dass mit der geplanten Inbetriebnahme oft später als geplant begonnen werden kann, da einige Komponenten längere Lieferzeiten als erwartet benötigen oder im Zuge der Montage festgestellt wird, dass teilweise grundle-

gende Änderungen der Mechanik nötig werden. Vor allem solche Redesigns der Anlage verzögern den Beginn der eigentlichen Programmierphase.
Aufgrund der zeitlichen Terminierung und festgesetzten Lieferzeitpunkten steht Inbetriebnehmern bzw. Programmieren somit meist weniger Zeit zur Verfügung als anfänglich geplant. Dass in solchen Fällen Einbußen in der Softwarequalität, bzw. dem Testen selbiger gemacht werden müssen, erscheint unvermeidbar.
Betrachtet man all diese Aspekte, wird der Mehrwert eines möglichst realen Modells der Maschine deutlich. Ein solches Modell ist an keinerlei Lieferzeiten gebunden und steht somit, zumindest in visueller Form, direkt nach der mechanischen Entwicklung zur Verfügung.
Legt man diesem visuellen Modell ein physikalisches Verhalten zu Grunde, wird es prinzipiell möglich, einige der oben genannten Arbeitsschritte vorzuziehen und auf diese Weise den eigentlich geplanten Workflow weitestgehend zu erhalten oder die Inbetriebnahmezeit sogar zu verkürzen. (vgl. [3], S. 1)

2 Problemstellung

Um dieses allgemeine Problem zu lösen, wurde das Softwarepaket *Virtuos* der Firma *ISG Industrielle Steuerungstechnik GmbH*[1] aus *Stuttgart* eingesetzt.
Gemeinsam mit dem Partner, der *IT-Engineering GmbH*[2], wurde auf praktische Weise, die durch die Software gegebenen Möglichkeiten mit Bezug auf die Echtzeitsimulation von Industrieanlagen untersucht.
Die zur Verfügung stehende Zeit wurde neben dem Erlangen der grundlegenden Techniken zur selbständigen Projektierung der Anlagen für eine umfassende Bewertung der Software und deren Integration in den gesamten Entwicklungsprozess einer Automatisierungsanlage genutzt. Hierbei wird das Hauptaugenmerk auf den Nutzen bei der letztlichen Softwareentwicklung auf der SPS gelegt.

Zu forderst gilt es folgende Fragestellungen zu klären:

- Kann die Integration einer solchen Simulation nahtlos in den bisherigen Entwicklungsprozess erfolgen? Und somit die Inbetriebnahmezeit einer Anlage reduzieren?
- Kann bei der Verwendung eines Modells zur Entwicklung der SPS-Software eine Steigerung der Softwarequalität gegenüber dem herkömmlichen Vorgehen erzielt werden?
- Wie real, bzw. wie detailliert kann ein Modell mit *Virtuos* erstellt werden? Wo liegen die Grenzen des Softwarepaketes?
- Welche Vorteile für die Entwickler und letztlich ein Unternehmen bietet die Verwendung eines Simulationsmodells im industriellen Umfeld während und auch nach der eigentlichen Inbetriebnahme?
- Entsteht letzten Endes auch ein finanzieller Nutzen für das einsetzende Unternehmen?

Um diese Fragen zu beantworten, gilt es eine real vorhandene Anlage mit Hilfe der *Virtuos*-Software nachzubilden, das Verhalten des Modells zu überprüfen und auf diese Weise die Verwendbarkeit des Softwarepaketes zur Echtzeitsimulation zu bewerten.

[1] Die Firma ISG bietet ein breites Softwarerepertoire in den Bereichen Steuerungs-, Antriebs- und Simulationstechnik an. Weitere Informationen unter: http://www.isg-stuttgart.de

[2] Die IT-Engineering GmbH versteht sich als individueller Softwareentwickler und realisiert auf diese Weise verschiedenste, kundenspezifische Projekte. Neben Individuallösungen wird auch ein hauseigenes MES-System angeboten. Weitere Informationen: http://www.ite-web.de/

Das nachzubildende System[3] ist bereits verkaufsreif und ist dementsprechend bereits über die Inbetriebnahmephase hinaus. Aus diesem Grund steht bereits ein fertiges Steuerungsprogramm zur Verfügung, welches die Anlage steuert. Daher eignet sich die zu betrachtende Anlage sehr gut für eine umfassende Machbarkeitsstudie. Sollten hier das konkrete Verhalten der modellierten Anlage, ohne Änderung des vorhandenen Steuerungsprogrammes, dem der realen Maschine entsprechen, können umfassende Rückschlüsse auf die Verwendung solcher Modelle während der Inbetriebnahme gezogen werden. (vgl. [1] S. 2)

3 Einführung in Virtuos und Systemaufbau

Um eine vollständige Simulation einer industriellen Maschine zu gewährleisten, werden mehrere Systemkomponenten gebraucht um das Simulationsmodell zu erstellen. Abbildung 1 zeigt eine Übersicht über die beteiligten Systeme und deren grundlegenden Zusammenhang.

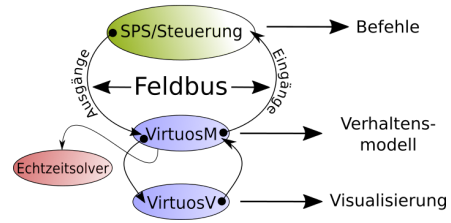

Abbildung 1: Grundlegender Systemaufbau des Simulationssystems

Um eine möglichst realitätsnahe Simulation zu gewährleisten, setzt ISG bei ihrem System auf eine sogenannte „Hardware-in-the-loop"-Anordnung (siehe [5]). Hierbei ist die real vorhandene Steuerung, welche auch bei der späteren Maschine zum Einsatz kommt, über ein physisches Feldbussystem verbunden. Nach diesem Prinzip tritt das Rechnersystem, auf welchem das *Virtuos*-Modell läuft, als aktiver Feldbusteilnehmer auf und kann auf diese Weise Ausgänge der SPS als Befehlssignale auffassen und die Eingänge der projektierten Steuerung nutzen, um die Antwort des Systems (z. B. Taster, Sensoren ect.) nachzubilden. (vgl. [2], S. 1 f)
Die eingesetzte Software besteht aus zwei unterschiedlichen, eigenständig laufenden Programmen, welchen jeweils folgende Aufgaben zugeschrieben sind:

- **VirtuosV**
  Diese Software ist für die Visualisierung des Modells zuständig. Hier wird das Modell in Form von CAD-Daten bereitgestellt und dem Benutzer angezeigt. An dieser Stelle wird definiert, welche grafischen Knoten sich bewegen und ob es sich dabei um eine translatorische oder eine rotatorische Bewegung handelt. Auf diese Weise können die unterteilten Knoten einzeln oder im Kollektiv bewegt werden.

- **VirtuosM**
  Das Modell, welches in *VirtuosM* hinterlegt wird, beschreibt das physikalische Verhalten der Anlage auf die Steuerungssignale. An dieser Stelle wird demnach die Reaktion der Maschine modelliert. Um dies zu erreichen, stehen an dieser Stelle einige physikalische Grundelemente wie pneumatische Zylinder zur Verfügung. Dies unterstützt die Integration solcher zeitlicher Modelle. Die Projektierung in *VirtuosM* erfolgt überwiegend in einem grafischen Programmierstil. Diese bausteinorientierte Definition des Modells erinnert an Simulink[4] und ermöglicht eine ähnliche Arbeitsweise.

Um den Zugriff der Ein- und Ausgangssignale für die *VirtuosM*-Software zu ermöglichen, kommt

[3] Aus Gründen der Geheimhaltung, ist es leider nicht möglich, genauere anlagenspezifische Details im Zuge dieser Ausarbeitung zu veröffentlichen.

[4] Weitere Informationen siehe: http://mathworks.com

auf dem Rechnersystem des Modells die Soft-SPS „TwinCat 3.1“ von Beckhoff[5] zum Einsatz. Durch diese Integration wird die direkte Teilnahme am Feldbus, und somit die Hardware-in-the-loop Simulation, möglich.

Neben all den Anbindungen benötigt das System zur Bereitstellung der Funktionalitäten noch den Echtzeitsolver. Dieser wird wahlweise direkt im Windows Betriebssystem implementiert oder bei Verwendung einer realen Steuerung direkt in der SoftSPS auf dem Simulationsrechner integriert. So wird eine harte Echtzeit der Simulationsberechnungen möglich.

In Kombination enthalten beide Softwareteile somit Werkzeuge zur Verbindung, Visualisierung sowie zur Modellierung einer Industrieanlage und bieten daher grundlegend die Möglichkeit, das in Abschnitt 2 gestellte Problem zu lösen.

4 Analyse der Möglichkeiten mit Virtuos

4.1 Projektierung mit Virtuos

Um das virtuelle Pendant einer Maschine nachzustellen, werden verschiedenste Elemente benötigt. Neben der Aktorik, welche sich meist aus pneumatischen oder hydraulischen Elementen und rotatorischen sowie translatorischen Antriebskomponenten zusammensetzt, muss auch die Sensorik bedacht werden.
Für den Bereich der Aktoren bietet *Virtuos* bereits eine Vielzahl verschiedener, vordefinierter Bausteine. Auf diese Weise wird es möglich, das Verhalten der gängigsten dieser Aktoren (z. B. bistabile, monostabile pneumatische Elemente oder sercosbasierende Antriebe) ohne großen Aufwand stabil in das physikalische Modell zu integrieren und mit dem Visualisierungsmodell zu verknüpfen. Abbildung 2 zeigt die beispielhafte Darstellung einer kompletten Integration eines bistabilen Zylinders in das VirtousM-Modell.

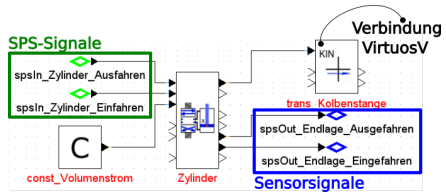

Abbildung 2: Beispielhafte Integration eines bistabilen Zylinders in das Maschinenmodell

Neben dem eigentlichen Signallauf kann durch Anpassung diverser Parameter und Ansteuern von zusätzlichen Eingängen (z. B. Zylinderdurchmesser, Kolbendurchmesser, Volumenstrom) weiteren Einfluss auf das zeitliche Verhalten des Modells genommen werden. Diese Einflussnahme verändert die Ansteuerung des, ebenfalls in Abbildung 2 zu sehenden, Kinematikbausteins, welcher das Visualisierungsmdoell mit dem Verhaltensmodell verknüpft und die Bewegungen an das 3D-Modell der Maschine überträgt.

Neben solchen Standardkomponenten kann es nötig werden, dass zusätzliche, seitens *Virtuos* nicht bereitgestellte, Bausteine benötigt werden. Hierzu bietet die Entwicklungsumgebung auch die Möglichkeit, Zustandsgraphen zu implementieren. Hierbei kommt die Programmiersprache **An**weisungs**l**iste (AWL) zum Einsatz. Diese Graphen können beispielsweise zur Definition der Zustandsautomaten für spezielle Frequenzumrichter genutzt werden.

Sollten auch diese Möglichkeiten nicht ausreichen, wird ein **S**oftware **D**evelopment **K**it (SDK) zur Verfügung gestellt. Dieses SDK ermöglicht eine direkte Integration von eigens definierten, C++-basierten Bausteinen in den Echtzeitsolver des Simulationssystems.
Diese hochsprachenbasierte Definition von einzelnen Teilmodellen kann daher für kompliziertere Rechenvorgänge genutzt werden, für welche der herkömmliche, grafische Ansatz eher ungeeignet scheint.

Neben den in obiger Abbildung 2 zu sehenden, direkt in den Baustein des Zylinders integrierten, Endlagensensoren können die Eingangssignale der SPS auf beliebe Art und Weise generiert werden. *Virtuos* bietet so zum Beispiel Positionsvergleiche der verschiedenen Positionsknoten an, um benötigte Sensorik nachzuempfinden.

4.2 Detaillierungsgrad

Da die Realität, beziehungsweise das reale, physikalische Verhalten einer komplexen Anlage an mehrdimensionale Freiheitsgrade gebunden ist, welche es bei einer allumfassenden Simulation zu berücksichtigen gilt, sind jeglichen bekannten Simulationen Grenzen auferlegt.
Vor allem, wenn der Fokus auf die virtuelle Inbetriebnahme von Industriemaschinen gelegt wird, gilt es im Vorfeld einen sinnvollen Detaillierungsgrad festzulegen. Auf Basis dieses **L**evel **o**f **d**etail (LOD) gilt es das Model zu entwickeln und die benötigten Funktionalitäten herzustellen. (vgl. [2], S. 1)

Beschränkt man sich auch hier auf den Bereich der virtuellen Inbetriebnahme, kann der LOD entsprechend niedrig gehalten werden. Da für die Programmierung des grundlegenden Steueurungsverhaltens auf die Ist-Zustände und Benutzereingaben der Maschine kein exaktes physikalisches Modell benötigt wird, kann hier viel Projektieraufwand eingespart werden. Auf diese Weise können sämtliche Bewegungsabläufe der Maschine ausreichend nachgestellt werden.

Innerhalb spezieller Anlagen kann es allerdings auch nötig werden, dass bestimmte Prozesse auf aktuelle Sensorinformationen zugreifen und diese den Prozess an sich beeinflussen. An dieser Stelle wird ein tiefergehenderes Modell mit entsprechend höherem LOD nötig.
Virtuos bietet in diesem Bereich, nicht zuletzt durch das zur Verfügung stehende C++-SDK, umfassende Möglichkeiten, die verschiedenen LODs innerhalb eines Models dynamisch zu gestalten und ermöglicht so einen nahezu beliebig zu definierenden LOD.

5 Einsatz des Modells

Steht ein virtuelles Abbild der betrachteten Maschine zur Verfügung, kann das Simulationssystem, gemäß Abbildung 1, aufgebaut werden und mit der Inbetriebnahme der Anlage begonnen werden.

Hierbei stellt das Modell sämtliches, mit Hilfe der in Abschnitt 4 beschriebenen Möglichkeiten, hinterlegtes Verhalten der realen Maschine zur Verfügung.
Nach dem bekannten Wechselwirkungsprinzip reagiert das Modell auf die Ausgangssignale der SPS mit dem hinterlegten, zeitlichen Verhalten und meldet mit Hilfe der SPS-Eingänge das benötigte Feedback der Maschine an die Steuerung und somit deren Programme.

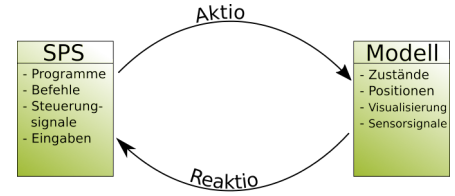

Abbildung 3: Actio und Reactio bei Verwendung der simulierten Anlage

Durch die direkte Anbindung der modellierten Anlage an den Feldbus erhält die SPS die simulierten Signalverläufe der Maschine direkt über die dortig hinterlegte Hardwarekonfiguration. Deshalb werden keinerlei Eingriffe in den bisherigen Entwicklungsprozess der Steuerungssoftware nötig.

Während das Modell, beziehungsweise dessen zeitliches Verhalten, im *VirtuosM*-Programmteil in Echtzeit berechnet wird, werden die durchzuführenden Aktionen

[5] Weitere Informationen siehe: http://www.beckhoff.de

direkt in der dreidimensionalen Umgebung von *VirtuosV* dem Benutzer visualisiert[6].

Auf diese Weise erhält der Entwickler direktes, visuelles Feedback des Maschinenerhaltens, welches durch die SPS-Routinen angestoßen wird und kann dieses direkt auf Richtigkeit überprüfen. Das Modell bietet somit die Möglichkeit, bereits vor abgeschlossener Montage Softwareroutinen zu entwickeln und diese, vor allem ohne reales Risiko eines Crashs der Maschine zu testen.

Das in *Virtuos* modellierte Modell bietet ebenfalls die Möglichkeit Bedienelemente direkt in der Simulationsumgebung zu erstellen. Dadurch ist gewährleistet, dass auch die vorgesehenen Benutzereingaben, wie durch Taster, Schalter und Sicherheitsmechanismen (zum Beispiel Nothalt), welche sich nicht direkt durch das physikalische Modell der Maschine generieren lassen, berücksichtigt werden können.

6 Bewertung der Simulation mit Virtuos

Im Zuge dieser Evaluation der Simulationssoftware *Virtuos* ist ein umfassendes Modell einer realen Anlage entstanden, welches, gesteuert durch die Steuerung und deren bereits vorhandenen Programmen, das Verhalten der zu Grunde gelegten Maschine originalgetreu widerspiegelt.
Um dieses Simulationsziel zu erreichen, wurden keinerlei Veränderungen an der SPS oder der Steuerungsoberfläche des Bedienpanels nötig. Deshalb lässt sich bereits jetzt zusammenfassend sagend, dass das verwendete Softwarekonstrukt grundsätzlich zur Durchführung einer virtuellen Inbetriebnahme geeignet ist.

Der grafische Projektierstil erschien für den überwiegenden Teil der Arbeiten sinnvoll. Dieser erreicht jedoch bei steigender Komplexität schnell die Grenzen der Übersichtichkeit, weshalb das C++-SDK als sinnvolle Erweiterung des Softwarepakets anzusehen ist.
Die dynamische Anpassung des LOD für die einzelnen Systembereiche (z. B. Bewegung des Modells oder Prozesse) erwies sich als probates Mittel zur Simulation des Modells.

Im Zuge der Modellentwicklung und den dadurch ermöglichten Testszenarien konnten bisher unentdeckte Softwarefehler in verschiedenen beteiligten Softwarekomponenten (z. B. HMI und SPS) aufgedeckt und teilweise auch bereits behoben werden. Hierfür zeigten sich vor allem die nun möglichen, unbeaufsichtigten Langzeittests, ohne zwingende Notwendigkeit einer vorhandenen Anlage, verantwortlich. Diese Tatsache lässt die Schlussfolgerung zu, dass das Vorhandensein eines virtuellen Modells durchaus zu einer Steigerung der Softwarequalität führen kann.
Hierfür zeigen sich neben der schnellen Durchführung und Wiederholbarkeit der Tests auch das relativierte Crashrisiko und nicht zuletzt die ruhige Büroumgebung, welche bei Inbetriebnahmen vor Ort nicht zwingend gegeben ist, verantwortlich.

Aufgrund der hohen, zeitlichen Aufwendungen, welche für die Evaluation der Software, sowie zum Erlangen der benötigten Grundfertigkeiten aufgebracht werden mussten, lässt sich der reell geleistete Modellierungsaufwand nur schätzen. Um ein grundlegendes, für Inbetriebnahmezwecke geeignetes Modell der behandelten Maschine zu erstellen, wird der zu leistende Zeitaufwand auf 100 Stunden geschätzt.
Je nach Komplexität der Anlagen ist dieser Richtwert nach oben bzw. nach unten zu korrigieren. Vor allem zusätzliche Sensorik oder Freiheitsgrade wirken sich auf die zu leistenden Modellierungsaufgaben aus.

[6] Ein Beispielvideo zur *Virtuos* Software findet sich z.B. unter: www.youtube.com/watch?v=Nui4CmkeqQ4

7 Ausblick

Neben der Verwendung solcher Modelle zur Verhaltenssimulation industrieller Anlagen bei der Inbetriebnahme können den generierten Modellen weitere Aufgaben zu Teil werden.
Nach der Definition des Steuerungsprogrammes bietet der Lebenszyklus einer solchen Anlage viele weitere Einsatzmöglichkeiten (siehe [4]). Hierbei ist vor allem an Problemanalysen von weit entfernten Maschinen, Machbarkeitstests bei Prozessanpassungen oder auch vertriebliche Dinge (z. B. Anlagenpräsentationen) zu denken.

Die zukünftige Arbeit mit den *Virtuos*-Modellen wird zeigen, ob neben der Inbetriebnahme der Maschinen auch solche weiteren Problemstellungen durch die hier evaluierten Möglichkeiten des Softwarepaketes *Virtuos* gelöst werden können.

Zum letztlichen finanziellen Nutzen durch den Einsatz solcher Simulationen kann keine zufriedenstellende und allgemeingültige Aussage getroffen werden. Es ist jedoch anzunehmen, dass ein früherer Beginn der eigentlichen Inbetriebnahme den geplanten Entwicklungsprozess, zumindest dessen zeitliche Abfolge im Positiven beeinflusst. Ob dieser jedoch den im Vorfeld zu leistenden Mehraufwand für die Modellierung der Anlage rechtfertigt, gilt es in einem umfassenden Feldversuch zu untersuchen.
Innerhalb dieses Versuches gilt es auch die Anpassbarkeit der Modelle genauer zu betrachten. Speziell im Bereich des Sondermaschinenbaus werden im laufenden Entwicklungsprozess oft vielerlei, teils gravierende Änderungen an dem mechanischen Aufbau der Maschine vorgenommen. Aus diesem Grunde sollte auch das virtuelle Modell einer solchen die Möglichkeit zur schnellen Anpassung an die neuen Umstände bieten.

Literatur

[1] M. Bergert, J. Kiefer, S. Höme, and C. Fedrowitz. Einsatz der virtuellen inbetriebnahme im automobilen karosserierohbau - ein erfahrungsbericht. *etz*, 9, 2010.

[2] A. Kufner, I. K. Haug, and I. P. Klemm. Modellierung von montagemaschinen für die hardware-in-the-loop-simulation. In *Paderborner Workshop Entwurf Mechatronischer Systeme (EMS), S*, pages 115–126, 2008.

[3] J. Mewes. Virtuelle inbetriebnahme mit realen automatisierungssystemen und virtuellen maschinen. *Hennigsdorf: Mewes*, 2005.

[4] U. Sendler. *Das PLM-Kompendium: Referenzbuch des Produkt-Lebenszyklus-Managements*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[5] G. Wünsch. *Methoden für die virtuelle Inbetriebnahme automatisierter Produktionssysteme*. Utz, 2008.

User Experience Design in der Medizintechnik - Einsatzmöglichkeiten von UX-Methoden innerhalb eines Entwicklungsprozesses von Medizinprodukten

Sebastian Hirth
Reutlingen University
Sebastian.Hirth@Student.Reutlingen-University.DE

Abstract
Zwar kann ohne die Einbindung des Endbenutzers ein normenkonformes Medizinprodukt entwickelt werden, jedoch können bei der Gestaltung von Bedienkonzepten der Benutzerschnittstelle grobe Designschwächen entstehen, welche zu einer unbefriedigenden User Experience beim Endanwender führen können. Die User Experience, welche ein Produkt bietet, stellt auch in diesem Bereich schon längst einen wichtigen Wettbewerbsfaktor dar. Daher soll innerhalb dieser Arbeit untersucht werden, welche Möglichkeiten bestehen, gängige Methoden des UX-Designs innerhalb eines Entwicklungsprozesses für Medizinprodukte anzuwenden.

Betreuer Hochschule: Prof. Dr. Burgert
Hochschule Reutlingen
Oliver.Burgert@Reutlingen-University.de
Betreuer Firma: Andreas Walden
Philips Medizinsysteme GmbH
Andreas.Walden@Philips.com
Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

Schlüsselwörter
Medical Device, Human Factors, User Experience, User Centred, Design

CR-Kategorien
H.5.2 [Information Interfaces and Presentation]: User Interfaces---User-centred design;

1 Einleitung
Dass Interaktive Systeme in allen Formen innerhalb von Krankenhäusern und Intensivstationen zum Einsatz kommen, gehört seit Jahren zum Standard. Ein gut durchdachtes Medizinprodukt kann dem Benutzer eine hohe Effizienz und Effektivität bei seiner Arbeit bieten. Ein schlechtes Produktdesign hingegen kann das Personal, welches mit dem Produkt interagiert durch unerwartetes Verhalten oder Unberechenbarkeit vor große Probleme stellen. Designentscheidungen, welche auf reinen Annahmen und unabhängig von *User Feedback* getroffen werden, führen daher in gewissen Bereichen möglicherweise zu medizinischen Vorfällen mit katastrophalen Folgen [1].

Im Gegensatz zu weiteren Disziplinen der Informationstechnik halten sich

interdisziplinäre Forschungsarbeiten zwischen dem Bereich der Medizintechnik und dem Bereich der Mensch-Computer-Interaktion und speziell des User Experience Designs (UXD) in Grenzen. Geht es um die Entwicklung von Medizinprodukten, so wurden in den letzten Jahren jedoch verstärkt Richtlinien und Standards eingeführt, welche eine Anwendung von benutzerzentrierten Methoden innerhalb des Entwicklungsprozesses bis zu einem gewissen Grad verordnen [2]. Werden diese Standards eingehalten, gilt ein Medizinprodukt als normenkonform und ist für den Markt zulässig. Dadurch sollen Funktionalität, Sicherheit und Gebrauchstauglichkeit sichergestellt werden. Der benutzerzentrierte Entwicklungsprozess im Bereich der Medizintechnik deckt jedoch nicht alle Vorgehensweisen des UXD ab [3]. Obwohl die UX eines Produktes auch in der Medizintechnik schon längst einen wichtigen Wettbewerbsfaktor darstellt [4], kommt diese Disziplin bei vielen Herstellern zu kurz. Dabei stellt die konsequente Einbindung des Benutzers und die Anwendung von benutzerzentrierten Methoden innerhalb des Entwicklungsprozesses viele Hersteller vor eine große Herausforderung [2]. Innerhalb dieser Ausarbeitung soll der praktische Standpunkt der Verwendung von UXD-Methoden innerhalb der Industrie untersucht werden. Neben der theoretischen Untersuchung aus der Herstellersicht wurde zudem ein normenkonformes Medizinprodukt auf Designschwächen, oder auch '*Design Fault*' untersucht, welche durch die Anwendung von UXD-Methoden frühzeitig im Entwicklungsprozess erkannt und somit vermieden werden können. Des Weiteren wurden zwei Forschungsarbeiten untersucht, welche sich mit der Frage beschäftigen, wie benutzerzentrierte Methoden innerhalb gängiger Vorgehensmodelle der Softwareentwicklung eingebunden werden können und ob diese Methoden im Kontext der Medizintechnik praxistauglich sind. Im ersten Teil der Ausarbeitung wird zunächst auf die gesetzlichen Regularien und die Anwendung von Usability-Methoden in der Medizintechnik eingegangen. Anschließend wird näher Erläutert was unter dem Begriff „User Experience“ innerhalb dieser Ausarbeitung zu verstehen ist und es folgt die Produktanalyse. Abschließend folgt ein Modell einer benutzerzentrierten Vorgehensweise im Entwicklungsprozess.

2 Usability Engineering in der Medizintechnik

Um die Sicherheit des Patienten und des Personals zu gewährleisten, unterliegen Medizinprodukte bestimmten Regularien. Die Richtlinien und Normen, welche bestimmte Anforderungen an ein Produkt stellen, variieren dabei deutlich. Die Tatsache, dass in vielen Ländern verschiedene Gesetzeslagen vorliegen, macht es den Herstellern schwer ein Produkt mit der Einhaltung geltender Gesetze und Richtlinien zu entwickeln. Die wichtigsten Regulationen stellen dabei die *EC Medical Device Directive (MDD)* und die *US Food and Drug Administration (FDA)*. Nur durch die gesetzlich vorgeschriebene Einhaltung dieser Richtlinien, kann ein Produkt in Europa (MDD) und den USA (FDA) auf den Markt gebracht werden [5]. Wie bei vielen weiteren Richtlinien liegt der Fokus jedoch hauptsächlich auf dem des *Risk Managements* und konzentriert sich somit auf den sicheren Gebrauch des Produktes.

Bei der Art der Umsetzung und der Gewährleistung der zusammenhängenden Anforderungen der Gebrauchstauglichkeit und Sicherheit eines Produktes gab es in den letzten Jahren jedoch Fortschritte und ein Umdenken.

“*The devices must be designed and manufactured in such a way that, when used under the conditions and for the purposes intended, they will not compromise the clinical condition or the safety of patients, or the safety of users* [6].”

Dieses Zitat, aus dem Standard 93/42/EEC (1993) der MDD, welches die Sicherheit eines Produktes anspricht, wurde im Jahr 2007 in einer modifizierten Version wie folgt abgeändert.

„Reducing, as far as possible, the risk of use error due to the ergonomic features of the device and the environment in which the device is intended to be used (design for patient safety), and consideration of the technical knowledge, experience, education and training and where applicable the medical and physical conditions of intended users (design for lay, professional, disabled or other users) [7].”

Bei dem Zitat der modifizierten Version aus dem Jahre 2007 wird gezielt auf den Endanwender, dessen Arbeitsumfeld und Eigenschaften eingegangen. Diese benutzerzentrierte Sicht war auch damals schon aus dem *Usability Engineering* bekannt. Innerhalb aktueller Standards ist dabei die Rede vom sogenannten *Human Factors Engineering*, welches ein benutzerzentriertes Vorgehen während der Entwicklung voraussetzt. Der von der MDD harmonisierte Standard IEC 62366 legt ebenfalls fest, dass die Hersteller von Medizinprodukten ausdrücklich auf die Usability ihrer Produkte eingehen müssen, bevor sie auf den europäischen Markt gelangen. Innerhalb des Standards wird die Bedeutung des *Usability Engineerings* und des *Human Factor Engineerings* als gleich angesehen [3]. Auch dieser Standard steht in enger Beziehung mit dem *Risk Management*.

“The purpose of the USABILITY ENGINEERING PROCESS, as described in this standard, is to provide SAFETY for the PATIENT, USER and others related to USABILITY. To achieve this purpose, the USABILITY ENGINEERING PROCESS mitigates RISK caused by USABILITY problems associated with CORRECT USE and USE ERRORS [3].”

Zwar nutzt der Standard IEC 62366 die Richtlinien der Mensch-Computer-Interaktion anhand der Norm EN ISO 9241, jedoch wird ebenfalls darauf hingewiesen, dass der Standard nicht sicherheitsbezogene und somit auch gesetzlich nicht verpflichtende Usability-Aktivitäten nicht komplett umfassen kann. Die FDA versucht an dieser Stelle jedoch Abhilfe zu schaffen, indem beispielsweise ein für alle Hersteller frei zugänglicher Leitfaden bereitgestellt wird [8]. Der Leitfaden umfasst dabei nichtbindende Empfehlungen an die Hersteller und zeigt Vorgehensweisen auf, welche über die gesetzlich verordneten Usability-Aktivitäten hinausgehen. Die Beschreibung einer UX, welche sich durch die Interaktion auf emotionaler Ebene äußert, wird anders als in der ISO Norm 9241 – 210 in keinem Fall beschrieben, was naheliegend ist. Die Leitfäden und Normen behandeln ausschließlich die pragmatische, funktionale und sicherheitsrelevante Seite eines Produktes. Die German UPA, der Berufsverband der Deutschen Usability und User Experience Professionals, hingegen nimmt das Thema der UX eines Produktes in ihren aktuellen Leitfaden „Usability in der Medizintechnik“ auf und beschreibt auch zahlreiche Methoden aus dem UXD [9].

2.1 Herstellersicht

Laut Martin & Barnett [10] bestehen drei Hauptfaktoren, welche die Hersteller von Medizinprodukten motivieren sollten, während der Produktentwicklung Nutzerforschung zu betrieben. Wie bereits beschrieben, ergibt sich die Hauptmotivation durch die gesetzlichen Gegebenheiten, welche die Anwendung von benutzerzentrierten Methoden verlangen. Nebenbei können Erkenntnisse aus der Nutzerforschung jedoch genutzt werden, um weitere Interessengruppen, wie Stakeholder und Gelbgeber zu überzeugen. Dies spielt besonders für kleine oder mittelständische Unternehmen eine Rolle, welche auf Finanzierungen innerhalb der Entwicklung angewiesen sind. Als dritten Punkt sehen Martin & Barnett vorhandene und umfangreiche Literatur, welche Methoden beschreibt, wie durch die Nutzerforschung Ziele und Bedürfnisse der Endanwender ermittelt werden können. Dadurch ergibt sich für den Hersteller die Möglichkeit, ein

Produkt gezielt seiner Zielgruppe anzupassen.

Die Realität sieht diesbezüglich anders aus. Money et. al führten 2010 eine Studie durch, in der elf Hersteller von interaktiven Medizinprodukten befragt wurden [2]. Das Ziel war es, herauszufinden, welche Rolle und welchen Stellenwert die Einbindung der Endanwender innerhalb des, vom Hersteller praktizierten Entwicklungsprozess hat. Money et. al. kamen dadurch zu folgender Erkenntnis:

„There is a lack of existing primary research that explores the challenges and benefits of involving users specifically within the medical device development process, particularly from a medical device manufacturer's perspective [2]*."*

Von den elf befragten Herstellern gab nur einer an, regelmäßig einen ordnungsgemäßen benutzerzentrierten Entwicklungsprozess durchzuführen, welcher sich durch alle Phasen der Produktentwicklung zieht. Viele Hersteller erkennen den Mehrwert eines benutzerzentrierten Vorgehens nicht, erfüllen jedoch formell die Richtlinien. Dabei erscheinen seit Jahren neue Studien, welche den Mehrwert eines benutzerzentrierten Vorgehens unterstreichen [11]. Der Umfang und die Funktionalität eines Produktes wird oftmals allein auf der Seite des Herstellers definiert, welcher durch die nötige Durchführung von Usability-Tests das Mindestmaß eines benutzerzentrierten Vorgehens erfüllt. Doch allein die Aufbringung der benötigten Ressourcen und des „*Know-hows*", um dieses Mindestmaß zu erreichen, stellt für viele eine große Herausforderung dar. Eine weitere Barriere entsteht auf der organisatorischen Seite. Laut Money et. al. haben viele Hersteller Probleme die Methoden des *Usability Engineerings* in ihr bestehendes Vorgehensmodell zu integrieren. Dies hat zur Folge, dass diese Aktivitäten oftmals an spezialisierte Agenturen und Firmen weitergegeben werden.

3 User Experience Design

Neben dem klassischen *Usability Engineering* in der Medizintechnik, welches sich mit den pragmatischen Eigenschaften eines Produkts auseinandersetzt, wird so gut wie nie auf die hedonischen Werte eingegangen. Laut Hassenzahl nehmen Menschen neben der Funktionalität auch weitere Produkteigenschaften auf einer emotionalen Ebene war, welche sich innerhalb des Produktcharakters widerspiegeln [12]. Durch UXD sollen neben pragmatischen Werten auch die hedonischen Werte eines Produktes hervorgehoben werden. UX ist dabei mehr als das ästhetische Erscheinungsbild eines Produktes oder Spaß und Vergnügen während der Interaktion. Dies sind nur Teilaspekte, welche zur UX beitragen [13]. Alle Merkmale und Eigenschaften eines Produktes können in die UX des Benutzers einfließen und den Benutzer je nach Erwartungshaltung prägen. Der Fokus liegt auf der Erfahrung, die durch das Produkt entsteht [14]. Durch die Bewertung während der Interaktion fällt die Erfahrung je nach Erfüllung oder Frustration positiv oder negativ aus. Darauf nehmen neben der Usability eines Produktes alle weiteren Faktoren Einfluss [15]. Im Jahr 2010 wurde die ISO 13407 durch die neu aufgesetzte ISO 9241 ersetzt. Unter Teil 210: 'Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme', wurde die UX aufgenommen. Im Standard wird UX wie folgt beschrieben:

„User experience is a consequence of the presentation, functionality, system performance, interactive behavior, and assistive capabilities of an interactive system, both hardware and software. It is also a consequence of the user's experiences, attitudes, skills, habits and personality [16]*."*

Im UX-Design wird zwar versucht eine Gesamterfahrung zu gestalten, jedoch gibt es keine Garantie darüber, ob der Benutzer diese auch so wahrnimmt. Die UX eines Produktes ist subjektiv, jedoch kann durch die Anwendung von UX-Design Methoden im Entwicklungsprozess ein Produkt entstehen, welches eine durchweg positive UX bietet. UX-Design beinhaltet alle Methoden eines benutzerzentrierten Entwicklungsprozesses und somit auch die des Usability Engineerings der Medizintechnik.

4 Vorgehensweise

Um Einsatzmöglichkeiten von UXD-Methoden bei der Entwicklung von Medizinprodukten zu untersuchen, wurde ein normenkonformes interaktives Medizinprodukt zur Messung von Vitalparametern analysiert. Das Produkt ist für den Markt in Europa und den USA zulässig. Das Produkt verfügt über einen resistiven Touchscreen, welcher als Interaktionschnittstelle dient. Anhand der Untersuchung sollte in Erfahrung gebracht werden, wie die Benutzerschnittstelle eines normenkonformen Medizinproduktes einer Untersuchung nach gängigen Heuristiken der Mensch-Computer-Interaktion standhält. Dabei ist nicht auf sicherheitsrelevante Aspekte eingegangen worden, da diese bereits durch Regularien zur Zulassung des Produktes beachtet wurden. Die Usability eines Produktes spielt bei der UX des Benutzers natürlich eine zentrale Rolle. Daher wurden Aspekte untersucht, welche eine reibungslose Interaktion und die Zufriedenheit des Benutzers ermöglichen sollen.

Als erster Schritt wurde die Informationsarchitektur analysiert. Das UXD bietet Methoden, um potentielle Endanwender bereits bei der Entwicklung des Aufbaus und der Struktur einer Benutzerschnittstelle einzubinden [13] [17] [18]. Durch '*Reverse Engineering*' wurde innerhalb der Arbeit die Informationsarchitektur der Benutzerschnittstelle nachgestellt. Dies sollte Verbesserungsmöglichkeiten durch die Anwendung von UXD-Methoden bezüglich der Informationsarchitektur aufzeigen. Anschließend wurde die Untersuchung anhand von Heuristiken fortgeführt. Die Ergebnisse dieser Untersuchung wurden dokumentiert und sollten als Indikator für Designschwächen dienen. Gleichzeitig wurden die gefundenen Probleme einem Ebenenmodell zugeordnet und passende UI-Design Patterns vorgeschlagen. Abschließend wurden UXD-Methoden des benutzerzentrierten Entwicklungsprozesses ermittelt, welche die Entstehung solcher Designfehler frühzeitig im Entwicklungsprozess vermeiden.

4.1 Heuristiken nach Nielsen

Eine bekannte Methode des Usability Engineerings ist die heuristische Evaluation. Dabei wird eine Ansammlung von Usability-Prinzipien herangezogen, um Probleme einer Benutzeroberfläche zu identifizieren [19]. Laut Nielsen steigt mit der Anzahl der Evaluatoren auch die Anzahl der gefunden Usability-Probleme, siehe Abbildung 1.

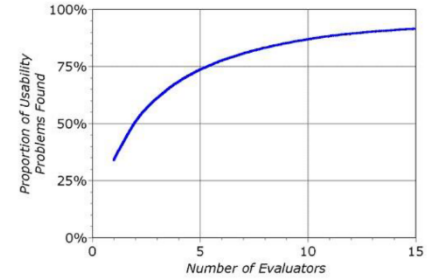

Abbildung 1: Abhängigkeit aufgedeckter Usability-Probleme und der Anzahl an Evaluatoren [20]**.**

Im Rahmen dieser Ausarbeitung wurde keine vollständige heuristische Evaluation durchgeführt. Ziel der Untersuchung war es nicht, alle vorhanden Usability-Probleme aufzudecken. Vielmehr ging es darum, eine Basis an Designschwächen zu erhalten, welche aufzeigt, dass alleine durch die Einhaltung gesetzlicher Vorschriften ohne die weiterführende Anwendung benutzerzentrierter Methoden gewisse Probleme bestehen.

4.2 Dialogprinzipien

Für die Untersuchung wurden ebenfalls die sieben Dialoggrundsätze für interaktive

Systeme nach ISO 9241-110 gewählt. Bei den Grundsätzen handelt es sich um Leitlinien, welche bei der Gestaltung von interaktiven Systemen beachtet werden sollten. Die Anwendung dieser Dialogprinzipien führt zu einer besseren Gebrauchstauglichkeit und Konsistenz der Benutzeroberfläche [21].

4.3 UI-Design Patterns

Als Quelle für die UI-Design Patterns wurde die Patternsammlung von Jenifer Tidwell aus dem Buch „Designing Interfaces“ verwendet [22]. Jenifer Tidwell publizierte mit „Common Ground“ 1998 die erste Zusammenstellung von Patterns für Benutzungsoberflächen [23]. Die in „Designing Interfaces“ vorgestellte Patternlanguage ist für eine Vielzahl von Interaktionsmöglichkeiten anwendbar und stellt eine umfassende Patternsammlung für den Bereich Mensch-Maschine-Interaktion bereit. Die Patternsammlung eignete sich gut für die Produktanalyse, da die verschiedenen Patterns je nach Einsatzmöglichkeit, wie beispielsweise der Informationsarchitektur oder der Navigation kategorisiert sind.

4.4 Ebenenmodelle

Um eine gewisse Taxonomie zu erhalten, wurden innerhalb der Arbeit verschiedene Ebenenmodelle untersucht. Durch die Betrachtung einer Benutzerschnittstelle anhand eines Ebenenmodells, können durch eine Kategorisierung der Interface Komponenten einzelnen Probleme besser zugeordnet, beschrieben und vor allem kommuniziert werden. Im Jahr 1992 führte IBM bereits ein solches Ebenenmodell für das Interface Design ein [24]. Im Gegensatz zu dem von IBM entwickelten Modell, welches aus den Ebenen Struktur, Verhalten und Präsentation bestand, griffen im Laufe der Zeit weitere Autoren diese Art der Beschreibung einer Benutzerschnittstelle auf.

4.4.1 Modell nach Garrett

Garrett sieht seinen Aufbau als eine Art konzeptuelles Gerüst für die Probleme der UX und liefert zusätzlich auch Werkzeuge, um diese zu lösen. Sein Ebenemodell umfasst alle Bestandteile einer Benutzerschnittstelle und lässt sich sehr gut innerhalb eines benutzerzentrierten Entwicklungsprozesses verwenden. Der Aufbau sollte von unten nach oben betrachtet werden. Angefangen auf der „Strategieebene“ werden bis zum „sensorischen Design“ alle Aspekte unter der Betrachtung des UX-Designs aufgeführt.

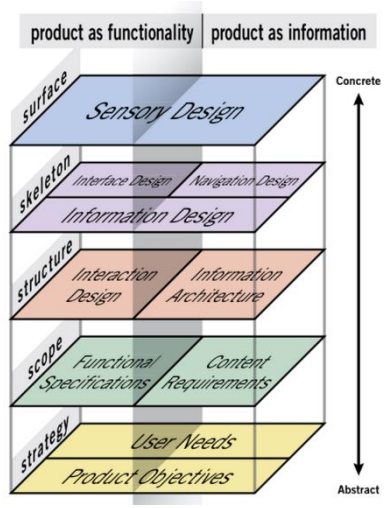

Abbildung 2: Die Elemente der User Experience nach Garrett [13]**.**

Die oberste Ebene stellt dabei die Summe aller Anwendungsebenen dar, mit welcher der Benutzer interagiert. Ein weiteres Merkmal besteht darin, dass Garrett die Inhalte der verschiedenen Ebenen in die Kategorien „Funktionalität“ und „Information“ unterteilt [13]. Das Modell von Garrett wurde für eine Zuordnung der einzelnen Designschwächen verwendet, da es die klaren Bestandteile einer Benutzeroberfläche aufzeigt und trotzdem genügend Komplexität bietet, um alle Verstöße gegen gewisse Heuristiken zuzuordnen. Für die Kategorisierung der aufgedeckten Designschwächen wurden die Elemente der Ebenen *Structure*, *Skeleton* und *Surface* verwendet.

5 Produktanalyse

Als erster Schritt erfolgte eine Untersuchung der Informationsarchitektur. Hierfür wurde ein Architekturdiagramm erstellt, welches die Struktur, Menüpunkte und Funktionen des Systems darstellt. Die anschließende Produktanalyse nach Heuristiken wurde von einem Evaluator durchgeführt. Dabei wurde die Benutzeroberfläche auf die Einhaltung der aus Tabelle 1 und Tabelle 2 zu entnehmenden Heuristiken untersucht.

5.1 Informationsarchitektur

Die komplexe Struktur und Menüführung der Applikation wurde innerhalb des Architekturdiagramms erfasst. Dabei wurden acht Ebenen definiert, wodurch eine Klassifizierung der Menüpunkte erfolgte (siehe Abbildung 3).

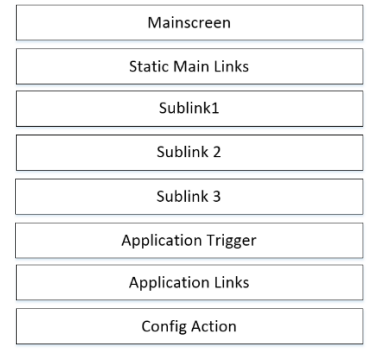

Abbildung 3: Ebenen des Architekturdiagramms.

Es konnte eine Anzahl von über 500 Menüpunkten identifiziert werden, wodurch deutlich wurde, dass das System eine organische Struktur besitzt und somit keinem regelmäßigen Muster folgt. Das Navigationsschema besteht aus Sprüngen zwischen den Ebenen. Es besteht keine Kategorisierung und Gruppierung zwischen den Menüpunkten des Hauptmenüs. Dadurch ergibt sich keine klare Informationsarchitektur, welche einer Struktur folgt. Die einzelnen Menüpunkte sind ohne Gliederung nach Bedarf verknüpft. Organische Strukturen liefern dem Benutzer keine Vorstellung, an welcher Stelle er sich innerhalb der Struktur befindet. Durch eine strukturierte Informationsarchitektur kann der Benutzer effizient und effektiv durch das System navigieren. Der Benutzer kann Menüpunkte und Inhalte ihrem Sinn nach erfassen und kognitiv besser verarbeiten. Dadurch sind Funktionen und Inhalte innerhalb des Systems einfacher aufzufinden [13]. Die untersuchte Benutzeroberfläche weißt diesbezüglich Schwächen auf.

5.2 Dialogprinzipien und Heuristiken

Um die einzelnen Probleme der Benutzerschnittstelle aufzeigen zu können, erfolgte eine Zuteilung der sieben Dialogprinzipien und den zehn Heuristiken nach Nielsen, siehe Tabelle 1 und 2.

Tabelle 1: Auflistung der Dialogprinzipien.

Dialogprinzipien	
DP1	Aufgabenangemessenheit
DP2	Selbstbeschreibungsfähigkeit
DP3	Steuerbarkeit
DP4	Erwartungskonformität
DP5	Fehlertoleranz
DP6	Individualisierbarkeit
DP7	Lernförderlichkeit

Tabelle 2: Auflistung der Heuristiken nach Nielsen.

Heuristiken nach Nielsen	
H1	Sichtbarkeit des Systemstatus
H2	Übereinstimmung zwischen System und realer Welt
H3	Benutzerkontrolle und Freiheit
H4	Konsistenz und Standards
H5	Fehlervermeidung
H6	Wiedererkennen vor Erinnern
H7	Flexibilität und effiziente Nutzung
H8	Ästhetisches und minimalistisches Design
H9	Hilfe beim Erkennen, Diagnostizieren und Beheben von Fehlern
H10	Hilfe und Dokumentation

Es wurden insgesamt 13 Designschwächen gefunden. Die Ergebnisse zeigen auf, dass

die Benutzeroberfläche insgesamt vier von sieben Dialogprinzipien (Abbildung 3) und sechs von zehn Heuristiken von Nielsen (Abbildung 4) nicht einhält.

Die Benutzeroberfläche verstößt in mehreren Fällen gegen die Dialogprinzipien

- Aufgabenangemessenheit
- Selbstbeschreibungsfähigkeit
- Steuerbarkeit
- Erwartungskonformität

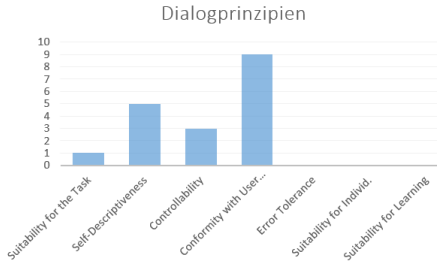

Abbildung 4: Auswertung nach den Dialogprinzipien.

Probleme bezüglich der Prinzipien Selbstbeschreibungsfähigkeit und Erwartungskonformität kamen sehr häufig vor. Zur Selbstbeschreibungsfähigkeit wurden fünf Probleme gefunden und zur Erwartungskonformität sogar neun. Dies spiegelt die Tatsache wider, dass die Benutzerschnittstelle an manchen Stellen keinen gängigen Konventionen folgt, wie sie im heutigen Interaktionsdesign üblich sind. Das mentale Modell eines Endanwenders beinhaltet diese Konventionen und erwartet ein bestimmtes Verhalten des Systems [25]. Hinzu kommt, dass die Benutzeroberfläche ihre Eigenheiten, welche nicht im mentalen Modell des Benutzers vorhanden sind, nicht optimal beschreibt und dem Benutzer aufzeigt.

Bei der Untersuchung anhand der Heuristiken von Nielsen wurden Probleme in Bezug auf die folgenden Regeln entdeckt:

- Sichtbarkeit des Systemstatus
- Übereinstimmung zwischen System und realer Welt
- Konsistenz und Standards
- Flexibilität und effiziente Nutzung
- Ästhetisches und minimalistisches Design

Abbildung 5: Auswertung nach Nielsen's Heuristiken.

Zur Heuristik Konsistenz und Standards wurden fünf Probleme gefunden. Die Benutzeroberfläche enthält teilweise inkonsistente Navigationsschemen oder es werden an Stellen, welche den Benutzer vor dieselbe Aufgabe bei der Ein- und Ausgabe von Daten stellt, verschiedene Interfaceelemente verwendet. Es fiel außerdem noch auf, dass sich das System trotz optisch identischer Dialoggestaltung nicht immer konsistent verhält. Somit verstößt das System teilweise gegen seine eigenen Konventionen. Weitere Verbesserungen könnten bei der Flexibilität und der effizienten Nutzung vorgenommen werden. Lange Navigationspfade und unnötige Klickwege wurden an drei Stellen festgestellt. Tabelle 3 listet nochmals alle Designschwächen auf.

Tabelle 3: Liste des Designschwächen.

Design Fault	Dialogprinzipien	Heuristiken
DF01	DP3, DP4	H4
DF02	DP2, DP4	H2, H4
DF03	DP1,DP2	H4
DF04	DP2,DP4	H4
DF05	DP2	H6
DF06	DP2, DP4	H4
DF07		H8
DF08	DP4	H8
DF09	DP4	H7
DF10	DP3, DP4	
DF11		H1
DF12	DP4	H7
DF13	DP3, DP4	H7

Durch die Untersuchung mit einem Evaluator konnten 13 Designschwächen erkannt werden. Insgesamt konnten dabei 28 Regelverstöße festgestellt werden. Dabei wurden 16 Dialogprinzipien und 12 Heuristiken von Nielsen verletzt. Laut Nielsen steht dieser Wert gerade mal für 35% aller vorhandenen Usability-Probleme [19]. Wie jedoch bereits beschrieben wurde, war es nicht das Ziel der Untersuchung alle Usability-Probleme aufzudecken. Anzumerken ist ebenso, dass es sich um ein sehr komplexes System mit einer umfangreichen Funktionalität handelt. Dabei kommt die untersuchte Benutzerschnittstelle Größtenteils den Dialogprinzipien und Heuristiken nach. Neben der Untersuchung wurden die ermittelten Designschwächen dem Ebenenmodell von Garrett zugeordnet. Die Designschwächen wurden nach den Teilebenen Navigationsdesign, Informationsarchitekur, Interfacedesign und dem sensorischen Design kategorisiert. Falls die Möglichkeit bestand, gewisse Probleme durch die Anwendung von UI-Designpatterns zu beheben, wurden diese ebenfalls dokumentiert. Tabelle 4 fasst die Erkenntnisse zusammen.

Tabelle 4: Zuordnung der Ebenen und Design-Patterns.

Design Fault	Ebene	UI-Pattern
DF01	Navigationsdesign	Pyramid
DF02	Informationsarchitektur	-
DF03	Navigationsdesign	Module Tabs, Titled Sections
DF04	Informationsarchitektur	Module Tabs, Titled Sections
DF05	Interfacedesign	-
DF06	Sensorisches Design	-
DF07	Navigationsdesign	Breadcrumbs
DF08	Navigationsdesign	-
DF09	Interfacedesign	-
DF10	Interfacedesign	Sortable Table
DF11	Interfacedesign	Progress Indicator
DF12	Interfacedesign	-
DF13	Navigationsdesign	-

Abschließend wurden aus einer Sammlung der German UPA [9] UXD-Methoden, welche die Problemstellungen der ermittelten Designschwächen behandeln zugeordnet. Methoden und Arbeitsweisen, welche Garrett auf den einzelnen Ebenen seines Modells vorschlägt wurden ebenfalls mit einbezogen. Die Methode der heuristischen Evaluation wurde innerhalb von Tabelle 4 nicht mit aufgeführt. Die ausgewählten Methoden stammen aus verschiedene Phasen des benutzerzentrierten Entwicklungsprozesses, behandeln jedoch die angegebene Designschwäche.

Tabelle 5: Liste der vorgeschlagenen Methoden.

Design Fault	Methoden
DF01	Mentale Modelle, Cognitive Walkthrough, Fokusgruppe, Usability-Test
DF02	Card-Sorting
DF03	Card-Sorting, Sequenzmodell
DF04	Card-Sorting, Architekturdiagramm, Sequenzmodell
DF05	Usability-Test, Cognitive Walkthrough, Styleguide, Mentale Modelle, Eye-Tracking
DF06	Styleguide
DF07	Sequenzmodell, Cognitive Walkthrough
DF08	Mentale Modelle, Styleguide
DF09	Eye Tracking, Mentale Modelle
DF10	Mentale Modelle, Usability-Test, Cognitive Walkthrough
DF11	Usability-Test, Cognitive Walkthrough
DF12	Mentale Modelle, Usability-Test, Cognitive Walkthrough
DF13	Sequenzmodell, Usability-Test, Cognitive Walkthrough

Die Mehrheit der in Tabelle 5 aufgeführten Designschwächen könnten durch iterative Usability-Tests während der Entwicklung behoben werden. Schwächen in der Informationsarchitektur und bei der Namensgebung von Menüpunkten könnten schon frühzeitig in der Konzeptphase ohne großen Kostenaufwand, beispielsweise durch die Card-Sorting Methode verbessert werden. Die Untersuchung konnte aufzeigen, dass bei einem normenkonformen interaktiven Medizinprodukt durchaus Designschwächen bestehen können. Ein benutzerzentrierter Entwicklungsprozess, welcher über die gesetzlichen Regularien hinaus geht und die Anwendung von UXD-

Methoden vorsieht, könnte das Benutzererlebnis in diesem Fall erheblich verbessern. Somit ist diese Art der Produktanalyse im Rahmen dieser Arbeit durchaus als sinnvoll zu betrachten, jedoch für weitere Verwendungszwecke nicht repräsentativ genug. Dies liegt an der Tatsache, dass die Evaluation auf den subjektiven Einschätzungen von nur einem Evaluator basieren. Des Weiteren erfolgte innerhalb der Untersuchung keine Gewichtung und Diskussion der identifizierten Probleme innerhalb einer Expertenrunde [19].

6 Entwicklungsprozess

Viele Hersteller von Medizinprodukten verwenden traditionelle Vorgehensmodelle bei der Entwicklung ihrer Produkte. Laut Money et. al. besteht ein Bedarf an Forschungen und Studien, wie sich ein benutzerzentrierter Entwicklungsprozess in der Praxis umsetzen lässt [2]. Viele Hersteller sparen an benutzerzentrierten Vorgehen, da die Integration innerhalb eines bestehenden Entwicklungsprozesses für sie einen erheblichen Mehraufwand bedeutet. Wie bereits unter Punkt 2.1 beschrieben, sind viele nicht bereit diesen Schritt zu gehen, da ihnen der Mehrwert einer kompletten Einbeziehung ihrer Endanwender nicht klar ist, oder sie glauben, unnötig viele Ressourcen aufwenden zu müssen. Laut Shluzas & Leifer, liegt die Problematik jedoch oft nicht an den Methoden selbst, sondern daran, wie und wann diese richtig eingesetzt werden [26]. Neben dem benutzerzentrierten Entwicklungsprozess, wie er im ISO Standard 9241-210 beschrieben wird [16], bietet die German UPA eine textuelle Beschreibung speziell für die Entwicklung von Medizinprodukten [9]. Auch eine Forschungsarbeit der *Multidisciplinary Assessment Of Technology Centre For Healthcare* (MATCH) aus dem Jahr 2009 beschäftigte sich mit den Fragen, wie ein benutzerzentrierter Entwicklungsprozess für Medizinprodukte gestaltet werden muss, um in jeder Phase die Möglichkeit zu haben, die Endanwender in die Entwicklung mit einzubeziehen [27]. Innerhalb dieser Arbeit wurden beide Beschreibungen untersucht. Um die UXD-Methoden aus Kapitel 5 innerhalb eines Entwicklungsprozesses unterzubringen, wurde ein Modell aus der textuellen Beschreibung der German UPA und dem Framework der MATCH abgeleitet (siehe Abbildung 6). Der Aufbau des Modells richtet sich dabei nach dem MATCH Framework, welcher jedoch die Inhalte der textuellen Beschreibung der German UPA enthält.

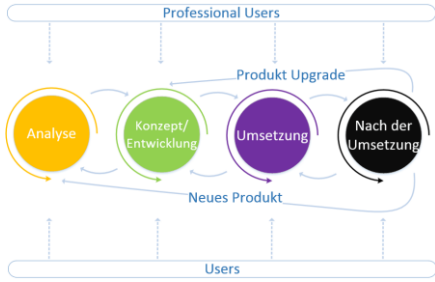

Abbildung 6: Kombination des MATCH Frameworks und dem des UPA Modells.

Der Leitfaden der German UPA und das MATCH Framework zeigen Wege auf, wie Hersteller von Medizinprodukten ihren Entwicklungsprozess anpassen könnten, um als Resultat neben einem normenkonformen Produkt eine noch höhere Gebrauchstauglichkeit und UX zu erreichen. Innerhalb der vier Phasen des Modells lässt sich eine Vielzahl von UXD-Methoden unter Einbeziehung der Endbenutzer anwenden. Unabhängig davon, ob es sich dabei um geschultes oder ungeschultes Personal handelt.

7 Fazit

Aus der Arbeit geht hervor, dass ein vollständiges benutzerzentriertes Vorgehen bei der Entwicklung von Medizinprodukten selten stattfindet. Die Hersteller kommen ihrer gesetzlichen Pflicht bei der Anwendung von benutzerzentrierten Methoden zwar formell nach, jedoch folgt nach der Pflicht selten die Kür. Zum Thema UX konnten im Kontext der Medizintechnik fast keine Studien oder weitere Forschungsarbeiten gefunden werden, jedoch umso mehr Informationen darüber, dass ein Bedarf an interdisziplinären Arbeiten besteht. Die durchgeführte Produktanalyse gab ebenfalls zu erkennen, dass eine Anwendung von UXD-Methoden bei der Entwicklung von Medizinprodukten eine weitere Steigerung der Gebrauchstauglichkeit zu Folge haben kann. Dies äußert sich im Idealfall durch eine erhöhte Zufriedenheit und eine positive UX beim Endanwender. Innerhalb eines benutzerzentrierten Entwicklungsprozesses, wie er von der German UPA oder dem MATCH Framework beschrieben wird, gibt es viele Einsatzmöglichkeiten solcher UXD-Methoden.

8 Literaturverzeichnis

[1] Y. Li, P. Oladimeji, C. Monroy, A. Cauchi, H. Thimbleby, D. Furniss, C. Vincent, and A. Blandford. Design of interactive medical devices: Feedback and its improvement. In *ITME 2011* - IEEE International Symposium on IT in Medicine and Education. 2011.

[2] A. G. Money, J. Barnett, J. Kuljis, M. P. Craven, J. L. Martin, and T. Young. The role of the user within the medical device design and development process: medical device manufacturers' perspectives. In BMC Medical Informatics and Decision Making. 2011.

[3] ISO/IEC 62366:2015, "Medical Devices - Application of Usability Engineering to Medical Devices," 2015.

[4] J. Lampkin and W. Buijs: Interaction Design and Medical Devices : Toward a Better User Experience. http://www.mdtmag.com/article/2014/07/interaction-design-and-medical-devices-toward-better-user-experience 2014. Accessed 25.10.2015

[5] J. L. Martin, B. J. Norris, E. Murphy, and J. A. Crowe. Medical device development: The challenge for ergonomics. In Applied Ergonomics Vol 39. S. 271–283. 2008.

[6] The European Parliament and the Council of the European Union. Medical Device Directive 93/42/EEC," *Off. J. Eur. Union*, 1993.

[7] The European Parliament and the Council of the European Union. Medical Device Directive 2007/47/EC. In Official Journal of the European Union. 2007.

[8] Food and Drug Administration. Draft Guidance for Industry: Applying Human factors and Usability Engineering to Optimize Medical Device Design. 2011.

[9] German UPA: Usability in der Medizintechnik. http://www.germanupa.de/data/mediapool/medizientechnik.pdf. 2015. Accessed 26.10.2015.

[10] J. L. Martin and J. Barnett. Integrating the results of user research into medical device development: insights from a case study. In BMC Medical Informatics and Decision Making Vol. 12 S. 74. 2012.

[11] J. L. Martin, D. J. Clark, S. P. Morgan, J. a. Crowe, and E. Murphy. A user-centred approach to requirements elicitation in medical device development: A case study from an industry perspective. In Applied Ergonimics Vol. 43, S. 184–190. 2012.

[12] M. Hassenzahl. The Thing And I: Understanding The Relationship Between User And Product. In M. A. Blythe Funology. From Usability to enjoyment. S. 31–42. 2003.

[13] J. J. Garret.The Elements of User Experience: User-Centred Design for the Web and Beyond. Pearson Education, Berkeley, 2011.

[14] V. Roto, E. L.-C. Law, A. Vermeeren, and J. Hoonhout. User Experience White Paper - Bringing clarity to the concept of user experience. 2011.

[15] E. Law, A. Vermeeren, M. Hassenzahl. Towards a UX Manifesto. In COST294-MAUSE affiliated workshop. 2007.

[16] EN ISO 9241-210:210. Ergonomics of human-system interaction Part 210: Human-centred design for interactive systems. 2010.

[17] A. Moed, M. Kuniavsky, and E. Goodman. Observing the User Experience. Morgan Kaufmann, Boston, 2012.

[18] J. Arango, P. Morville, and L. Rosenfeld. Information Architecture. O'reilly Media Inc. Sebastopol, 2015.

[19] J. Nielsen, Usability Inspection Methods. Wiley & Sons, New York, 1994.

[20] J. Nielsen, *Usability Engineering*. AP Professional, Cambridge, 1993.

[21] EN ISO 9241-110:2006. Ergonomics of human-system interaction Part 110: Dialogue principles. 2006.

[22] J. Tidwell. *Designing Interfaces*. O'Reilly Media Inc. Sebastapol, 2011.

[23] J. Tidwell, Common Ground - A Patternlanguage for Human-Computer Interface Design. http://www.mit.edu/~jtidwell/common_ground.html 1999. Accessed 20.10.2015.

[24] IBM. Object-Oriented Interface Design - IBM Common User Access Guidelines. New York, 1992.

[25] R. Hartson and P. Pyla. *The UX Book*. Morgan Kaufmann, Boston, 2012.

[26] L. M. Aquino Shluzas and L. J. Leifer. The insight-value-perception (iVP) model for user-centered design. In Technovation, Vol. 34, S. 649–662, 2014.

[27] S. G. S. Shah, I. Robinson, and S. AlShawi. Developing medical device technologies from users' perspectives: a theoretical framework for involving users in the development process. In International Journal of Technology Assessment in Health Care, Vol. 25, S. 514–521, 2009.

Onboarding in Business Software: Unterstützung von Erstnutzern

Simone Liegl
Reutlingen University
Simone.Liegl@Student.Reutlingen-University.DE

Abstract

Das Schwierige ist es nicht, einen Nutzer zum Anmelden in einer neuen Anwendung zu bewegen, sondern ihn zu halten. Eine Möglichkeit zur Unterstützung ist ein Onboarding, welches dem Nutzer hilfreiche Informationen gibt. Ziel ist, die First Use Experience zu verbessern, die sich aus der Nutzerzufriedenheit und Effektivität ergibt. Am Beispiel der Ressourcenmanagement-Software *Meisterplan* wird untersucht, ob ein Onboarding ersteres erhöht. Zur Messung der allgemeinen Akzeptanz wird ein Firstclick-Test durchgeführt. Ein Nutzertest mit Fragebogen und Eyetracker dient zur Bewertung der User Experience. Die Evaluation zeigt, dass die Abbruchrate des Onboardings 39% beträgt und außerdem mit steigendem Alter abnimmt. Eine geführte Unterstützung wirkt sich weiterhin positiv auf die Nutzerzufriedenheit und Effektivität von Erstnutzern aus.

Schlüsselwörter

Onboarding, First Use Experience, Akzeptanz, Nutzerzufriedenheit, Effektivität.

Betreuer Hochschule: Prof. Dr. Gabriela Tullius
Hochschule Reutlingen
Gabriela.Tullius@Reutlingen-University.de
Betreuer Firma: Stefan Schneider
itdesign GmbH
Stefan.Schneider@itdesign.de

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

CR-Kategorien

H.5.2 **Information Interfaces and Presentation**: User Interfaces.

1 Motivation

„You don't get a second chance to make a first impression" [8]. Diese Aussage lässt sich unter anderem auf den Bereich der Software-Entwicklung anwenden. Das Schwierige ist nicht, einen Nutzer zum Anmelden zu einer neuen Anwendung zu bewegen, sondern ihn bei der Nutzung zu unterstützen und dadurch als Kunden zu halten. Dieser soll sich nicht verloren oder überwältigt fühlen, sondern wissen, wie er effizient mit dem neuen Produkt arbeiten kann [8]. Der Nutzer muss also nicht nur in seinen Fähigkeiten sondern auch in seiner Motivation gefördert werden, um das gewünschte Verhalten zu erzielen [7]. Dies stellt in Business-Software wie Projektmanagement-Anwendungen eine Herausforderung dar, da aufgrund der vielfältigen Einsatzmöglichkeiten und Schnittstellen eine hohe kognitive Last entstehen kann; der Nutzer fühlt sich überfordert. Eine Möglichkeit, diese Last zu verringern, besteht in der Unterstützung durch einen persönlichen Mentor, was jedoch durch hohe Personalkosten und zeitliche Rahmenbedingungen beschränkt ist. Durch den Einsatz von Automatisierung kann der Nutzer stattdessen orts- und zeitunabhängig durch eine geführte Einleitung unterstützt werden, welche ihm die wichtigsten Funktionen der Software näher bringt. Dennoch ist hier der Lernerfolg nur schwer zu überprüfen. *„[T]here*

is sometimes a mismatch between how an educator wants to teach and what is represented on the interface by the instructional designers. Such mismatches affect the learner's experience and his motivation [...]" [13].

An dieser Stelle setzt die folgende Arbeit an, um den Bereich des Onboardings[1] zu evaluieren. Ein solches zielt auf die Verbesserung der First Use Experience ab, was der Periode zwischen dem ersten Ein- und Ausloggen entspricht. Hierbei müssen die Bedürfnisse eines Erstnutzers erfüllt werden. Nach Alderfer lässt sich nach drei solcher Bedürfnisse unterscheiden: Existenz, Beziehung und Wachstum. Für die vorliegende Arbeit ist insbesondere letztes relevant. Dieses umfasst die Wachstums- und Selbsterfüllungsbedürfnisse einer Person im Hinblick auf Selbstverwirklichung und Produktivität [1]. *„Growth needs include all the needs which involve a person making creative or productive effects on himself and the environment. Satisfaction of growth needs comes from a person engaging problems which call upon him to utilize his capacities fully and may include requiring him to develop additional capacities"* [1]. Eine Anwendung des Bedürfnisses auf ein Onboarding bedeutet damit die einfache und effektive Nutzung der Software, welche zu einer hohen Nutzerzufriedenheit führt. In diesem Zusammenhang stellt sich die Frage, ob sich die First Use Experience durch die Integration eines Onboardings erhöhen lässt und der Einstieg in komplexe Business-Software dadurch erleichtert werden kann. Voraussetzung hierfür ist, dass dieses vom Nutzer angenommen und nicht abgebrochen wird.

Ziel der Arbeit ist es, diese Fragen aufzugreifen und zu beantworten. Anhand der Ressourcenmanagement-Software *Meisterplan* wird evaluiert, wie groß die Akzeptanz eines Onboardings ist und ob sich dieses auf die First Use Experience auswirkt. Die Forschungsfrage lautet damit: Verbessert ein Onboarding die First Use Experience in Business-Software?

2 Wissenschaftlicher Hintergrund

Basis für die weitere Arbeit bildet der wissenschaftliche Hintergrund. Dieses Kapitel ist eine zielgerichtete Auswahl der Aspekte, die für die Ableitung der Hypothesen relevant sind.

2.1 Informationsverarbeitung

Die Informationsaufnahme des Menschen ist aufgrund von verschiedenen Prozessen in der Informationsverarbeitung limitiert; das Kurzzeitgedächtnis des Menschen verfügt über einen begrenzten Speicher. Information wird Kategorien, sogenannte Chunks, eingeteilt [11]. Diese sind *„groups of items that have been collected together and treated as a single unit"*. [12] Je mehr von diesen verarbeitet werden müssen, desto längere Reaktionszeiten und mehr Fehler entstehen. In der Literatur werden verschiedene Maximalmengen an Chunks genannt, die zwischen vier und sieben variieren [11]. Nach Miller (1956) liegt die Kapazität des Kurzzeitgedächtnisses bei sieben plus minus zwei. Neuere Untersuchungen hingegen zeigen, dass die Kapazität mit einer Zahl von vier plus minus eins geringer ist. Diese wird unter anderem durch den zeitlichen Abstand der Information oder die Aufgabenstellung beeinflusst [12]. Es ergeben sich resultierend folgende Argumente, welche für die bewusste Limitierung von Information bei deren Vermittlung sprechen: *„Weakness - which suggests that larger limits would overload mental mechanism, and Strength - which argues the advantage of a limited and well structured representation"* [11].

Die Informationsverarbeitung basiert auf einem bottom-up und top-down Modell der Aufmerksamkeit [5]. Diese Idee eines Zwei-Komponenten-Modells geht auf William James zurück und besagt, dass die Aufmerksamkeit durch zwei Aspekte gesteuert wird. Die bottom-up Komponente behandelt eine vom frontalen Gehirnlappen kontrollierte Aufmerksamkeit, die automatisch im Bereich von 25 bis 50 ms pro Objekt stattfindet und durch Salienz[2] gesteuert wird. Dem steht die top-down Komponente gegenüber, die aufgabenabhängig ist, sodass hierfür ein freiwilliger Aufwand geschehen muss. Beide können gleichzeitig arbeiten [9]. Für die vorliegende Arbeit ist insbesondere die aufgabenbasierte Aufmerksamkeit von Interesse, da diese im Onboarding[3] gezielt auf bestimmte Bereiche der Arbeit gelenkt wird.

2.2 E-Learning

Da Onboarding eine Form des E-Learnings ist, sind einige Aspekte daraus für die vorliegende Arbeit von Interesse.

Unter E-Learning wird das *„Lehren und Lernen mittels verschiedener elektronischer Medien"* [16] verstanden. Der auf Erfahrung basierende Lernzyklus beginnt mit einer neuen, konkreten Erfahrung. In der zweiten Phase werden vergangene Erlebnisse herangezogen, um die neuen aus verschiedenen Perspektiven zu betrachten. Danach erfolgt eine Konzeptualisierung, in der die Person Theorien und Lösungen für das Problem findet. Darauf schließt sich der Kreis durch ein aktives Handeln, wobei die Theorien getestet werden [2]. Durch eine Unterstützung in diesem Prozess kann die Menge des Wissens erhöht werden [13].

Die Menge und Art des Wissens, die gelernt werden kann, wird nicht nur durch Prozesse der Informationsverarbeitung, sondern durch das Individuum bestimmt. Neben der unterschiedlichen Ausprägung von Fähigkeiten [10] spielen auch Lernstile eine Rolle.

Kolb unterscheidet vier verschiedene Stile: Divergieren, Konvergieren, Assimilieren und Akkomodieren. Menschen vom ersten Typ sind sehr gut im Kombinieren von konkreten Erfahrungen und reflektierenden Beobachtungen, wobei die Information stets aus verschiedenen Ansichtspunkten betrachtet wird. Diesem Lerntyp steht der dominante Konvergierer entgegen, welcher abstrakte Konzepte in aktive Experimente transferiert, um die Theorie in der Praxis zu testen. Menschen vom Lerntyp Assimilieren arbeiten mit reflektierten Beobachtungen und setzen dabei abstrakte Konzepte ein [2]. Ihre Stärke liegt in der Bildung von theoretischen Modellen, wobei weniger die praktische Umsetzung von Ideen von Interesse ist. Akkomodierer kombinieren konkrete Erfahrungen mit aktiven Experimentieren. Ihr Fokus liegt im Ausführen und Testen von Plänen und Ideen, wobei diese auch verworfen werden, falls sie nicht funktionieren [10].

2.3 Onboarding

Onboarding unterstützt den bereits erläuterten Lernprozess durch eine geführte Tour in einem digitalen Produkt.

2.3.1 Definition

Es beschreibt den Kennenlern- Prozess in einer für den Nutzer neuen Anwendung. Dies umfasst Themenbereiche wie Registrierung, Aufnahme sowie Verstehen der Grundstruktur und Funktionalität [15]. *„ By providing onboarding mechanism, users will be enabled to smoothly pass into the efficient usage of the digital product. […] It can be defined as the slow increasement of the system's complexity, delivering positive reinforcement to avoid early fails and get to know the users"* [15].

Der Begriff stammt ursprünglich aus dem Bereich der Human Resources und bezeichnet dort den Prozess der Aufnahme, Assimilierung und Beschleunigung der Prozesse von neuen Mitarbeitern im Unternehmen [8].

[1] Da die deutsche Übersetzung ‚Einarbeitung' die Betonung auf den initialen Lernprozess legt und dabei das ständige Weiterentwickeln zum Experten außen vor lässt, wird in folgender Arbeit der englische Begriff verwendet, um die Breite des Begriffs zu verdeutlichen.

[2] Präaufmerksamer Prozess [1]

[3] Siehe Abschnitt 2.3.4

2.3.2 Formen

Ein Onboarding kann auf verschiedene Arten umgesetzt werden. Es zielt darauf ab, ein gewünschtes Verhalten zu erreichen. Hierzu muss der Nutzer über die Fähigkeit und Motivation verfügen, das Verhalten durchzuführen. Des Weiteren muss ein aktiver Reiz existieren, der zum richtigen Zeitpunkt erkennbar sein muss und in Verbindung mit dem Zielverhalten stehen muss. [7] Grundsätzlich lässt sich zwischen aktiven und passiven Formen unterscheiden. *„A passive tour consists of information about provided features. The navigation within the tour is independent from the fact if the users accessed this specific feature"* [15]. Dem steht die aktive Tour gegenüber, bei welcher der Nutzer eine Aufgabe lösen muss, um zum nächsten Schritt zu gelangen. Auch eine Mischform der beiden ist möglich [15]. Die aktive Tour spricht insbesondere die Lerntypen Akkomodierer und Konvergierer an.

Es existieren verschiedene Ansätze, nach denen ein Onboarding umgesetzt werden kann. Der des *Joyriding* ist eine passive Form, zeigt den Nutzer gezielt bestimmte Elemente und führt dazu, dass er sich nicht verloren fühlt [3]. Diese Art spricht Menschen vom Lerntyp Divergieren an. Beispiele hierfür sind eine Willkommensmail oder ein Tutorial [15]. Dem steht der aktive *Tu-etwas*-Ansatz entgegen, bei welchem der Nutzer dazu aufgefordert wird, sich aktiv in der Anwendung zu bewegen [3]. Dies kann durch ein Inline-Tutorial oder der Zuschaltung von neuen Funktionen mit fortlaufender Nutzung umgesetzt werden [15]. Der *Setup*- Ansatz geht davon aus, dass die Anwendung selbst einfach zu verstehen ist und die Einrichtung des Accounts das eigentliche Problem darstellt [3]. Alle vorgestellten Formen können sich miteinander vermischen, um auf die Anforderungen der Nutzer einzugehen.

2.3.3 Onboarding in Meisterplan

Meisterplan ist eine Software für Ressourcen- und Portfoliomanagement, welche von der itdesign GmbH entwickelt und vertrieben wird. Die Anwendung ermöglicht es, Projekte und darauf arbeitende Personen zu verwalten. Des Weiteren lassen sich verschiedene Planungsszenarien simulieren. Die Anwendung bildet alle Änderungen in Echtzeit ab. Durch die vielfältigen Einsatzmöglichkeiten steigt der Komplexitätsgrad in der Benutzung je nach Rolle des Benutzers. Dies impliziert die Notwendigkeit einer Unterstützung, um dem Nutzer die Möglichkeiten der Software aufzuzeigen.

Um einen potentiellen Kunden bei der ersten Nutzung zu unterstützen, ist in *Meisterplan* eine Onboarding-Tour integriert. Diese besteht aus zwei Formen: zum einen werden regelmäßig Emails versendet. Bei dem ersten Einloggen sieht der Nutzer zum anderen eine Inline-Tour, siehe Abbildung 1.

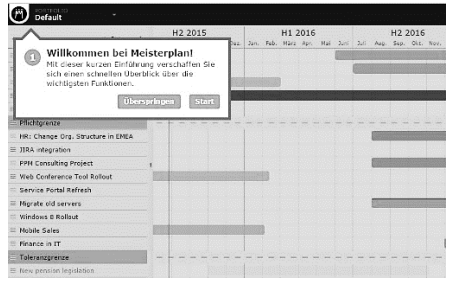

Abbildung 1: Onboarding in Meisterplan

Es gilt, den Nutzer einerseits zu motivieren und andererseits seine Fähigkeiten zu erhöhen. Durch einen aktiven Reiz anhand der Schritte wird er weiterhin gereizt. Die umgesetzte Tour ist eine Mischung einer aktiven und passiven Tour und erklärt dem Nutzer die drei Hauptansichten und die für den Anfang wichtigsten Funktionen. Die einzelnen Schritte selbst basieren auf dem Prinzip von Coachmarks, bei welchem an spezifischen Elementen Informationen gezeigt werden. Um zum jeweils nächsten Schritt zu gelangen, bedient sich die Tour zweier Möglichkeiten: einem „Weiter"-Button und einer Interaktion durch den Nutzer. Der Nutzer erhält damit einerseits auf dem klassischen Weg Informationen und muss diese andererseits interaktiv anwenden.

3 Hypothesen

Basierend auf der Forschungsfrage und begründet durch den wissenschaftlichen Hintergrund werden nun Hypothesen erläutert, welche am Beispiel von *Meisterplan* auf ihren Wahrheitsgehalt überprüft werden. Ziel ist, die Nullhypothese zu entkräften: Das Onboarding hat keine Auswirkung auf die First Use Experience in Business-Software.

3.1 Akzeptanz

Eine gute Akzeptanz ist Voraussetzung zur Messung der User Experience. Studien aus dem E-Learning-Bereich sprechen von Abbruchraten von bis zu 70%, welche in der langen Dauer, Ineffektivität oder schlechter Usability begründet liegen [13]. Renz et al. sprechen sogar von 75,7% Abbruchrate bei einer Onboarding-Tour in MOOCs[4] [15].

Weitere Erkenntnisse aus der Literatur zeigen, dass das Alter ein ausschlaggebender Faktor für die Akzeptanz ist. Begründet liegt dies in der Tatsache, dass sich Menschen in jungem Alter neues Wissen schneller und umfangreicher aneignen als ältere. *„Die für viele Kinder und Jugendliche selbstverständliche Auseinandersetzung mit neuen Medientypen und Medienformen, und der damit einhergehende Aufbau der Medienkompetenz, finden dabei fast ausnahmslos in der Form des Informellen Lernens statt"* [5]. Daraus resultiert die Vermutung, dass die Abbruchrate beim Onboarding von jüngeren Nutzern höher ist als von älteren.

Aus diesem Grund besagt die erste Hypothese, dass sich die Akzeptanz des Onboardings mit steigendem Alter erhöht. Um den Wahrheitsgehalt der Hypothese zu ermitteln, werden zwei Variablen verwendet: das Alter und die Abbruchrate.

3.2 Nutzerzufriedenheit

Wie bereits erläutert, ist die Nutzerzufriedenheit ein Faktor, der sich auf die User Experience auswirkt. Tullis und Albert nennen die Zufriedenheit als wichtige Metrik für die Messung von User Experience. Sie beschreiben Zufriedenheit als *„the degree to which the user was happy with his or her experience while performing the task"* [17].

Durch den hohen Komplexitätsgrad und die Menge an dargestellten Informationen in *Meisterplan* ist zu erwarten, dass Erstnutzer, welche die Software ohne jegliche Hinweise oder Erläuterungen kennen lernen, aufgrund der hohen kognitiven Last ein Gefühl von Verlorenheit empfinden. Im Gegenteil bedeutet dies, dass ein Onboarding den Nutzer unterstützt. Das Gefühl der Verlorenheit hängt mit dem Wohlfühlfaktor zusammen: je mehr sich ein Nutzer in Meisterplan zurecht findet, desto wohler fühlt er sich.

Zum Onboarding-Prozess gehört das Anwenden des Gelernten, wie in Abschnitt 2.2 deutlich wurde. Um zu überprüfen, ob der Nutzer dazu in der Lage ist, muss dieser im Experiment Aufgaben lösen. Hierbei werden das Verständnis der getätigten Aktion und die Zufriedenheit mit der eigenen Leistung gemessen. Für diese beiden Variablen ist zu erwarten, dass sich ein Nutzer, der von der Onboarding-Tour durch die Anwendung geführt wurde, zufriedener mit seiner Leistung ist und zudem bessere Ergebnisse erzielt.

Zusammenfassend besagt Hypothese 2, dass die subjektive Nutzerzufriedenheit der Onboarding-Testgruppe besser ist als die der Kontrollgruppe.

3.3 Effektivität

Um die Effektivität zu testen, müssen die Probanden im Test Aufgaben lösen. Tullis und Albert nennen den Aufgabenerfolg als gängige Metrik der User Experience [17]. *„A UX metrics reveal something about the interaction between the user and the product: some aspect of effectiveness (being able to complete a task)"* [17].

Es ist davon auszugehen, dass ein Nutzer durch das Onboarding einen Überblick über die wichtigsten Funktionen erhält und die Aufgaben damit besser löst. Für die Gruppe

[4] Massive Open Online Course

des Selfexploring hingegen kann es passieren, dass sich der Nutzer zu detailliert auf einige Bereiche konzentriert und damit der Überblick fehlt. Aus diesem Grund wird die Hypothese aufgestellt, dass sich das Onboarding positiv auf die Effektivität auswirkt.

Für die vorliegende Arbeit wird die Messung der Effektivität binär umgesetzt, eine Aufgabe kann also nur falsch oder richtig bearbeitet werden. Die Aufgaben decken verschiedene Bereiche von Meisterplan ab und wenden einerseits vermitteltes Wissen aus dem Onboarding an, verlangen aber auch Transfer und Entdecken von neuen Funktionen. Insgesamt sind es fünf Aufgaben, welche sich in elf Unteraufgaben aufteilen. Die Effektivität ergibt sich aus dem Durchschnitt der Summe aller Aufgaben eines Teilnehmers.

4 Experiment

Das Experiment wird mit einer Experimental- und Vergleichsgruppe durchgeführt. Es ist eine benutzerzentrierte Evaluation, bei der das tatsächliche Nutzungsverhalten und die subjektive Einstellung der Benutzer getestet werden. Es werden nicht nur subjektive Einstellungen, sondern auch Beobachtungsdaten erhoben.

4.1 Methodik

Für die Evaluierung des Onboardings in *Meisterplan* wird auf verschiedene Methoden zurückgegriffen.

Eine objektive Messung des Nutzungsverhaltens kann durch die Methode des Firstclick-Tests erreicht werden. *„In Firstclick-Tests werden Nutzer aufgefordert, den Bereich eines angezeigten Screenshots durch Anklicken zu markieren, hinter welchem sie die Lösung einer bestimmten Fragestellung vermuten“* [6]. Der Vorteil dieser Methode ist, dass er sowohl online mit anonymen Nutzern als auch face-to-face durchgeführt werden kann. Das Ergebnis, also die Summe aller Firstclicks der Nutzer, wird in der Regel als Heatmap dargestellt, welcher die Klick-Häufigkeit in den einzelnen Bereichen angibt [6]. Der Firstclick-Test wird verwendet, um die erste Hypothese zu überprüfen. Hierbei wird das Online-Tool von Usabilityhub[5] eingesetzt.

Eine zweite Methode stellt das Eyetracking dar. *„Eyetracking is simply following the trail of where a person is looking. […] According to the mind-eye hypothesis, people are usually thinking about what they are looking at. They do not always totally understand or engage with it, but if there are looking, they are usually paying attention, especially when concentrating on a particular task”* [14]. Mit der Auswertung der Eyetracking-Daten können verschiedene Dinge erfasst werden: Subjektives wie Zufriedenheit aber auch Objektives wie Effektivität [5]. Durch dessen Einsatz ist es also möglich, herauszufinden, wieso bestimmte Probleme existieren und nicht nur zu wissen, dass sie bestehen.

Um die Nutzerzufriedenheit zu testen, wird dem Nutzer nach Lösung der Aufgaben ein Fragebogen vorgelegt. Dafür wird die Online-Plattform SoSci Survey[6] eingesetzt. Dabei handelt es sich um einen standardisierten Fragebogen. In diesem werden zum einen die subjektive Nutzerzufriedenheit und zum anderen persönliche Daten gemessen. Die Variablen werden durch eine Likert-Skala erhoben.

4.2 Aufbau und Durchführung

Die Stichprobe umfasst 231 Nutzer, wobei 204 davon ausschließlich am Firstclick-Test teilnahmen. Dieser wurde aus Gründen der Reichweite auf Englisch durchgeführt. Beim Nutzertest hingegen konnten die Teilnehmer selbst entscheiden, ob sie Deutsch oder Englisch bevorzugen. Zur Aufzeichnung der Augenbewegungen wurde ein Tobii X2-30 Eyetracker mit einer Abtastrate von 30 Hz verwendet, welcher mit einem Laptop und zusätzlichem Bildschirm verbunden ist.

Bevor er am Test teilnimmt, wird jeder Nutzer in eine der beiden Gruppen – Onboarding oder Selfexploring - eingeordnet. Nach der Begrüßung durch den Moderator wird der Nutzer aufgefordert, den Einführungstext aus dem Fragebogen durchzulesen. Hierdurch wird garantiert, dass alle Teilnehmer über den gleichen Informationsstand verfügen. Nachdem der Eyetracker auf einen Nutzer kalibriert wurde, beginnt das Experiment mit dem Firstclick-Test. Der darauffolgende Test besteht je nach Gruppe in der freien oder geführten Einarbeitung, dem Lösen der Aufgaben und dem Beantworten des Fragebogens.

Die Verteilung der beiden Gruppen ist mit Stichproben-Größen von 14 für das Onboarding und 13 für die Kontrollgruppe gleich verteilt. Alterstechnisch nimmt die Gruppe zwischen 21 und 25 mit 27% den größten Anteil ein; die Verteilung erstreckt sich von 18 bis 65 Jahre. Die Variablen Alter und Geschlecht weisen keine Normalverteilung auf.

5 Evaluierung

Im Folgenden werden anhand der erhobenen Daten die einzelnen Alternativhypothesen auf ihren Wahrheitsgehalt überprüft.

5.1 Akzeptanz

Hypothese 1 behandelt die Akzeptanz des Onboardings. Die Abbruchrate beträgt insgesamt 39,9% und liegt damit unter den Werten von anderen Studien. Bei Teilnehmern des Nutzertests fällt auf, dass deren Abbruchrate lediglich 7,4% beträgt. Eine Analyse der Eyetracker-Daten zeigt, dass tatsächlich 22,2% abbrechen wollten, da ihr Blick deutlich länger auf dem Abbruch-Button verweilt als auf dem des Starts. Begründet werden kann dies durch den Hawthorne-Effekt, nach dem sich Nutzer der Beobachtungs- und Prüfungssituation bewusst sind und deshalb unterschiedlich agieren als sonst.

Abbildung 2 bildet die Abhängigkeit der Abbruchrate vom Alter ab. Es ist ersichtlich, dass die Abbruchrate bei steigendem Alter prozentual sinkt. Ausgenommen davon ist die Altersgruppe von 46 bis 50. Dies kann

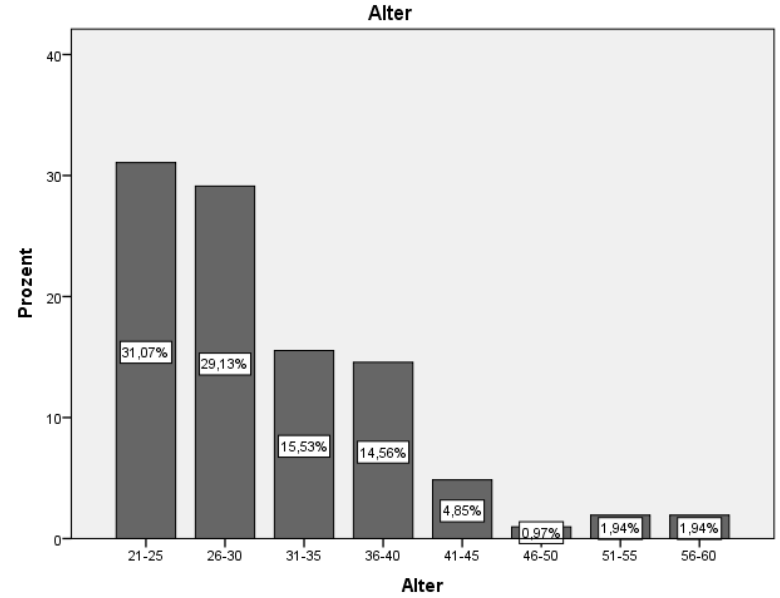

Abbildung 2: Abbruchrate nach Alter

[5] https://usabilityhub.com

[6] https://www.soscisurvey.de/

durch die geringe Repräsentation dieser Altersgruppe begründet werden. Insgesamt sinkt die Abbruchrate von 31,07% in der Altersgruppe von 21 bis 25 auf 1,94% in der Altersgruppe 56 bis 60. Untersuchungen nach Signifikanz[7] auf einen Zusammenhang dieser Variablen ergeben einen Wert von 10,1%. Damit ist das Signifikanzniveau von 5% nicht erfüllt, dennoch kann an dieser Stelle von einer Tendenz gesprochen werden.

5.2 *Nutzerzufriedenheit*

Hypothese 2 besagt, dass die subjektive Nutzerzufriedenheit der Onboarding-Testgruppe besser ist als die der Kontrollgruppe. Die folgende Auswertung der Daten bezieht sich auf die 27 Teilnehmer des Nutzertests. Die subjektive Nutzerzufriedenheit setzt sich aus vier Variablen zusammen: Zufriedenheit mit der eigenen Leistung, Verständnis der getätigten Aktionen, Wohlfühlen und Orientierungslosigkeit während der Benutzung, wobei alle Werte nach der Likert-Skala in fünf Ausprägungen bewertet werden.

Bei der Auswertung der Zufriedenheit mit der eigenen Leistung ist ersichtlich, dass die Gruppe mit freier Einarbeitung bessere Werte erzielt als die mit Onboarding. Da diese Aussage jedoch mit einem Wert von 0,5 nicht signifikant ist, sind die Ergebnisse kritisch zu betrachten. Des Weiteren fühlen sich Nutzer mit freier Einarbeitung bei der Benutzung von *Meisterplan* wohler als die Onboarding-Gruppe, bei welcher die Werte breiter und negativer gestreut sind. Allerdings gibt es auch zu dieser Aussage keine Bestätigung durch Signifikanz.

Beim Verständnis der getätigten Aktionen weist die Onboarding-Gruppe ein besseres Verständnis auf als die Kontrollgruppe[8]. Bei ersteren erstreckt es sich lediglich über die Werte „mittelmäßig“ bis „sehr gut“, in der Kontrollgruppe hingegen existieren alle Ausprägungen. Dies hängt möglicherweise damit zusammen, dass den Nutzern im Onboarding der Grundaufbau von *Meisterplan* näher gebracht wird und ihnen dadurch grundlegende Funktionsweisen klarer sind. Nutzer mit freiem Einarbeiten hingegen konzentrieren sich auf zu viele Details, wie Filter- und Einstellungsmöglichkeiten, wie die Auswertung der Eyetracker-Daten zeigt.

Sehr deutlich ist der Unterschied in der Orientierungslosigkeit. Nutzer mit dem Onboarding geben an, sich signifikant[9] weniger orientierungslos zu fühlen als solche mit freier Einarbeitung.

Aufgrund von fehlender Signifikanz ergibt sich die Nutzerzufriedenheit lediglich aus den Variablen Verständnis und Orientierungslosigkeit. Abbildung 3 zeigt, dass die Verteilung bei der Onboarding-Gruppe in einem positiveren Feld liegt als bei der Gruppe mit freier Einarbeitung. Dennoch ist die Verteilung in letzterer breiter gestreut. Insgesamt lässt sich festhalten, dass die Nutzerzufriedenheit, welche auf den Variablen „Verständnis der getätigten Aufgaben“ und „Orientierungslosigkeit während der Benutzung“ basiert, besser ist als in der Gruppe mit freier Einarbeitung.

Bei Analyse dieser Aussage hinsichtlich einer Signifikanz mit dem U-Test nach Mann-Whitney ergibt sich ein Wert von 0,02. Damit kann Hypothese 2 unter der Beachtung von zwei anstatt vier Variablen als bestätigt angesehen werden.

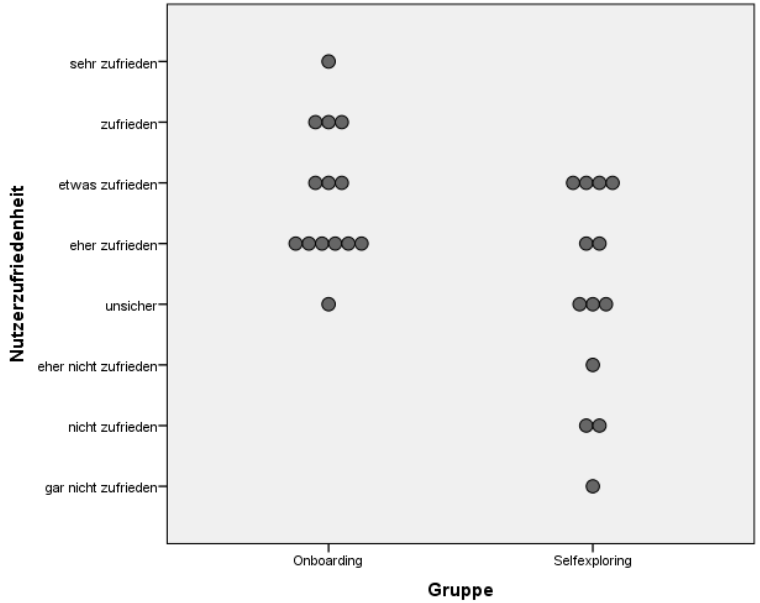

Abbildung 3: Nutzerzufriedenheit

5.3 *Effektivität*

Hypothese 3 besagt, dass die Effektivität in der Onboarding-Testgruppe besser ist als in der Kontrollgruppe.

Eine Auswertung ergibt bei lediglich drei der elf Unteraufgaben eine signifikante Aussage:

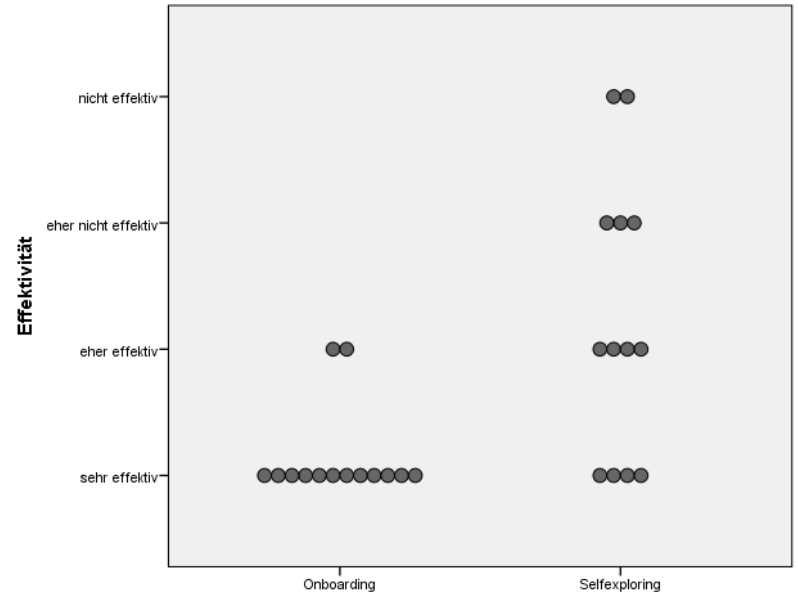

Abbildung 4: Effektivität

[7] Signifikanz-Test mit Spearman-Rho, bedingt durch die fehlende Normalverteilung

[8] Signifikanzwert von 0,11

[9] Signifikanzwert von 0,003

dem Verschieben eines Projekts[10], dem Hinzufügen einer Rolle[11] und der Vergabe von Kapazität[12] an diese. Aus diesem Grund ergibt sich die Effektivität aus eben genannten Werten. Die anderen Aufgaben weisen zum großen Teil darauf hin, dass die Testgruppe mit geführter Einarbeitung bessere Ergebnisse erzielt.

In Abbildung 4 ist sichtbar, dass die Verteilung in der Onboarding-Gruppe weniger breit gestreut und im besseren Teil angesiedelt ist. Während dort 44,4% die Aufgaben sehr effektiv lösen, sind es in der Gruppe mit freier Einarbeitung lediglich 14,81%. Des Weiteren ist bei letzterer die Effektivität geringer, da auch schlechte Werte wie „eher nicht effektiv“ und „nicht effektiv“ existieren. Der Signifikanz-Test zeigt, dass die Onboarding-Gruppe mit einer Wahrscheinlichkeit von 99,7% effektiver arbeitet. Unter Beachtung dieser drei Variablen kann Hypothese 3 somit als bestätigt angesehen werden.

5.4 Weitere Ergebnisse

Neben der Untersuchung der Hypothesen auf ihren Wahrheitsgehalt wurden die Daten auf weitere eindeutige Ergebnisse untersucht. So fällt auf, dass Nutzerzufriedenheit und Effektivität korrelieren: je effektiver der Nutzer arbeitet, desto zufriedener ist er mit der eigenen Leistung und umgekehrt. Eine Messung der Irrtumswahrscheinlichkeit zeigt, dass diese Aussage mit einem Wert von 0,005 sehr signifikant ist.

Des Weiteren korreliert die Variable „Wohlfühlen während der Benutzung“ linear mit allen anderen, welche die Nutzerzufriedenheit repräsentieren: dem Verständnis, der Orientierungslosigkeit und der Zufriedenheit. Die ersten beiden weisen ebenfalls einen Zusammenhang auf: je orientierungsloser die Nutzer bei der Benutzung sind, desto weniger Verständnis haben sie über die getätigten Aktionen und umgekehrt.

6 Fazit und Ausblick

Basierend auf der Beantwortung der drei Hypothesen lassen sich Rückschlüsse für die Forschungsfrage ziehen.

Da die Abbruchrate der Onboarding-Tour in *Meisterplan* mit 39,9% vergleichsweise gering ist, lässt sich von einer guten Akzeptanz sprechen, welche mit steigendem Alter zunimmt. Sowohl die Nutzerzufriedenheit als auch die Effektivität ist bei Nutzern, die das Onboarding durchlaufen, größer als bei solchen mit freier Einarbeitung. Da bei der Auswertung dieser beiden Hypothesen aufgrund von mangelnder Signifikanz nicht alle Variablen berücksichtigt werden können, sind die Ergebnisse noch weiter zu evaluieren. Dennoch lässt sich mit aktuellem Stand von klaren Aussagen sprechen. Rückschließend bedeutet dies, dass die Nullhypothese entkräftet werden kann. Gleichzeitig lässt sich die Forschungsfrage bejahen: Ein Onboarding wirkt sich positiv auf die First Use Experience in Business-Software aus.

Nach Cook reicht dennoch ein initiales Onboarding nicht aus [3], da der Komplexitätsgrad einer Anwendung mit tiefergehender Bedienung steigt. Stattdessen muss ein solches auch während der Benutzung weiter geführt werden, wobei an dieser Stelle ein richtiger Grad der Menge an Unterstützung gefunden werden muss. Ein Onboarding muss nicht zwingend eine Inline-Tour sein, stattdessen ist auch eine Kombination von verschiedenen Formen denkbar. Interessant ist es des Weiteren, eine solche Unterstützung auf einen Benutzer einerseits als Rolle und andererseits als Individuum anzupassen: wer ist der Nutzer, welche Rolle nimmt er ein, wofür möchte er *Meisterplan* nutzen und wo liegen seine fachlichen Schwerpunkte?

Die Beantwortung dieser Fragen muss im Rahmen von weiteren Arbeiten erfolgen, um den Onboarding-Prozess für ein digitales Produkt weiter zu verbessern.

[10] Signifikanz von 0,01

[11] Signifikanz von 0,06

[12] Signifikanz von 0,08

7 Referenzen

[1] Alderfer, C. P. An empirical test of a new theory of human needs. *Organizational Bevahior and Human Perfomance* 4, 2, S.142–175. 1969.

[2] Chen, C. C., Jones, K. T., and Moreland, K. Differences in Learning Styles. *CPA Journal* 84, 8, S.46–51. 2014.

[3] Cook, M. UX Flows: *How to Turn Onboarding into an Amazing First Date with Your User.* 2014. http://www.dtelepathy.com/blog/design/ux-flows-onboarding. Accessed 14 July 2015.

[4] Dittler, U. E-Learning. Lernen, Wissen und Bildung auf dem Weg in die Postmedialität. In *E-Learning. Einsatzkonzepte und Erfolgsfaktoren des Lernens mit interaktiven Medien*, U. Dittler, Ed. Informatik 10-2012. Oldenbourg Wiss.-Verl., München, S.1–27. 2011.

[5] Duchowski, A. T. *Eye Tracking Methodology. Theory and Practice*. Springer Verlag, London. 2007.

[6] Eck, K., Heuwing, B., and Womser-Hacker, C. Eignen sich Tree-Tests und Firstclick-Tests für die nutzerzentrierte Evaluierung der Informationsarchitektur? *Tagungsband der 13. Internationalen Symposiums für Informationswissenschaft*, S.286–297. 2013.

[7] Fogg, B. A behavior model for persuasive design. *Proceedings of the 4th International Conference on Persuasive Technology.* 2009.

[8] Hess, W. *Onboarding: Designing Welcoming First Experiences.* http://uxmag.com/articles/onboarding-designing-welcoming-first-experiences. 2010. Accessed 17 July 2015.

[9] Itti, L. and Koch, C. Computational modelling of visual attention. *Nature* reviews. Neuroscience 2, 3, S.194–203. 2001.

[10] Kolb, D. A. Learning Styles and Disciplinary Differences. In *The modern American College. Responding to the New Realities of Diverse Students and a Changing Society,* A. W. Chickering, Ed. Jossey-Bass, San Francisco, S.232–255. 1981.

[11] Mandler, G. The limit of mental structures. *The Journal of general psychology* 140, 4, S.243–250. 2013.

[12] Mathy, F. and Feldman, J. What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition* 122, 3, S.346–362. 2012.

[13] Minocha, S. and Sharp, H. Learner-Centered Design and Evaluation of Web-Based E-Learning Environments. *The 7th HCI Educators Workshop: Effective Teaching and Training in HCI*. 2004.

[14] Nielsen, J. and Pernice, K. *Eyetracking Web Usability*. New Riders, Berkeley. 2010.

[15] Renz, J., Staubitz, T., Pollack, J., and Meinel, C. Improving the Onboarding User Experience in MOOCS. 2014.

[16] Rey, G. D. *E-Learning und Multimedia*. http://www.elearning-psychologie.de/elearning_multimedia.html. 2015. Accessed 29 July 2015.

[17] Tullis, T. and Albert, B. *Measuring the user experience. Collecting analyzing and presenting usability metrics*. Elsevier, Amsterdam. 2013.

[18] Ware, C. *Information visualization. Perception for design.* Interactive technologies. Morgan Kaufmann, Boston. 2012.

Annotation medizinischer Textcorpora als Grundlage für Textminingverfahren*

Lasse Naumann
Reutlingen University
Lasse.Naumann@Student.Reutlingen-University.DE

Abstract

Nachfolgende Ausarbeitung befasst sich mit einem aktuellen Forschungsprojekt des Reutlingen Research Institutes. In Zusammenarbeit mit der Augenklinik der Universität Tübingen soll eine Lösung erarbeitet werden, die anhand gegebener Befundungsinformationen eines Patienten die entsprechende ICD-Codierung zur normierten Beschreibung der Diagnosen empfiehlt. Für die Umsetzung der Lösung wird ein Werkzeug benötigt, mit dessen Hilfe sich die Grundwahrheit für das Training eines Merkmalsextraktors erzeugen lässt.

Schlüsselwörter

Textmining, ICD Klassifikation, Text Annotation

CR-Kategorien

I.2.7 [**Natural Language Processing**]: Text analysis

1 Einleitung

Die medizinische Falldokumentation bietet einen bisher nur schwer greifbaren Reichtum an Wissen. Die Extraktion eben dieses Wissens und die computergestützte Verarbeitung und Analyse der das Wissen repräsentierenden Informationen ermöglichen einen immensen diagnostischen Mehrwert, vereinfachte Prozesse bei der Abrechnung der erbrachten Leistungen und nicht zuletzt eine qualitativ höhere Patientenversorgung.
So könnte in Zukunft bei Befundung und Diagnose nun nicht mehr nur Kompetenz und Erfahrung des behandelnden Arztes und seiner direkten Kollegenschaft von Bedeutung für den Behandlungserfolg sein, sondern der Patientenbefund könnte mit zahlreichen Befunden mit einer ähnlichen Symptomatik auf identische Merkmale verglichen werden, um daraus einen diagnostischen Mehrwert zu erhalten.
Um das in der klinischen Dokumentation inhärente Wissen computergestützt nutzen zu können, muss dieses allerdings erst in eine verwertbare Form aufbereitet werden.
Eine Lösung hierzu sind Systeme, die die patientenspezifische Falldokumentation auswerten und inhaltliche Merkmale auf Nomenklaturen wie ICD[1] (International Statistical Classification of Diseases and Related Health Problems), oder SNOMED[2] (Systematized Nomenclature of Medicine), abbilden. Der ICD stellt die internationale Standardklassifikation zur Diagnosestellung dar, wie sie auch in deutschen Krankenhäusern Verwendung findet. Durch die so ermöglichte formale Normierung individueller pathologischer Befunde, dient sie auch als Basis zur Verrechnung der erbrachten Leistungen mit den Krankenkassen.

*

Betreuer Hochschule: Prof. Dr. Christian Thies, Hochschule Reutlingen, Christian.Thies@Reutlingen-University.DE
Betreuer Firma: Dr. Lucien Clin, Reutlingen Research Institute, Lucien.Clin@Reutlingen-University.DE

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

[1] http://www.who.int/classifications/icd/en/
[2] http://www.ihtsdo.org/snomed-ct

Nach der Befundung eines Patienten und der entsprechenden Diagnosestellung, muss der Diagnose und eventuellen Prozeduren durch den Arzt, oder einen Dokumentar, der entsprechende ICD-Code zugeordnet werden.
Die aktuelle Version des ICD enthält weit über 13.000 Diagnosen und 3.500 Prozeduren. Auch wenn in den einzelnen Fachbereichen die potentielle Anzahl an verwendbaren Codes weitaus kleiner ist, so ist diese Zuordnung dennoch keine triviale Aufgabe und kann, abhängig von der Seltenheit der Diagnose, einige Zeit benötigen.
Genannte Systeme bringen hier eine erhebliche Erleichterung der dokumentarischen Arbeit. Besonders von Bedeutung für den Erfolg solcher Systeme sind alle Vorverarbeitungsschritte, da sie ausschlaggebend für die Qualität der grundlegenden Daten sind [1]. Demnach ist die Aufbereitung der Datenbasis, beispielsweise in Form der Annotation der zu kodierenden Informationen (Erkrankungen, Behandlungsmethoden, etc.), von grundlegender Bedeutung [9].

Das Forschungsprojekt befasst sich mit dieser Thematik. Die klinische Patientendokumentation soll durch ein Textminingverfahren analysiert werden. Dabei soll erörtert werde, inwiefern man durch Natural language Processing (NLP) anhand der in einem Befund enthaltenen Informationen auf gestellte Diagnosen schließen kann.
Ähnlich dem Vorgehen in [10] soll so die Möglichkeit gegeben werden, anhand bereits gestellter Befunde und deren Diagnosen, Diagnoseempfehlungen zu aktuellen Befunden zu stellen. Diese Empfehlungen können einerseits bei schwierigen pathologischen Gegebenheiten eine Hilfe sein, andererseits erleichtern sie die Codierung des Befundes in ICD.
Bevor dies allerdings erfolgen kann, wird ein Werkzeug benötigt, mit dessen Hilfe die Befunde zur Erschaffung einer Grundwahrheit annotiert werden können.

Diese Ausarbeitung beschreibt die grundsätzlichen Überlegungen bei der Erarbeitung dieses Werkzeugs. Kapitel 2 erläutert bereits vorhandene Ansätze, die eine vergleichbare Problematik adressieren. In Kapitel 3 wird die konkrete Problemstellung dargelegt und nachfolgend in Kapitel 4 bereits vorgestellte Ansätze auf ihre Anwendbarkeit bezüglich der Problemstellung überprüft. Kapitel 5 stellt die, anhand der gewonnenen Kenntnisse, entwickelte Methode dar, zu welcher in Kapitel 6 erste Ergebnisse präsentiert werden. Abschließend bietet Kapitel 7 einen Ausblick auf das weitere Vorgehen.

2 Verwandte Arbeiten

Es existieren bereits verschiedenste Ansätze für Systeme, die sich mit der Zuordnung von Elementen der natürlichen Sprache auf standardisierte Kataloge und Ontologien befassen.
Für die systemgestützte Extraktion von ICD-Codes aus der klinischen Dokumentation koexistieren nach [9] zwei vorherrschende Ansätze: Ein regel- und wissensbasierter und ein statistischer Ansatz, beide mit ihren Vor- und Nachteilen. Der regelbasierte Ansatz gilt als zeitaufwändig, da er die Formalisierung linguistischer Regeln zur Beschreibung des Domänenwissens für die Assoziation von Text zu Krankheiten umfasst. Diese Regeln müssen durch Experten der entsprechenden Domäne erstellt werden [6]. Der statistische Ansatz im Gegensatz erfordert im Allgemeinen einen geringeren Aufwand in der Vorverarbeitung der Corpora, benötigt aber einen sehr großen Datenbestand. Anhand linguistischer Merkmale (Features) und annotierter Trainingsdaten kann das System lernen.
Die logische Konsequenz hieraus ist, diese beiden Ansätze zu kombinieren, um sogenannte hybride Systeme zu erhalten [6].
Chiaravalloti et al. präsentieren hierzu einen probabilistischen Ansatz durch ein System, welches einen medizinischen Text analysiert, der Diagnosen oder medizinische Prozeduren beschreibt. Anhand der inhärenten Informationen der Texte werden diejenigen ICD-9-CM-Codes bereitstellt, die am wahrscheinlichsten mit dem Text zu assoziieren sind. Das System basiert auf Methoden der Textverarbeitung zur Extraktion der Informationen, einer Wissensbasis und einer Mustererkennung zum Abgleich dieser mit dem zu analysierenden Textcorpus [4].
Fette et al. stellen einen Ansatz vor, um einen annotierten Textcorpus für das Training eines automatischen Labelers für unstrukturierte medizinische Texte zu erzeugen [5]. Die Idee des Labelers besteht darin, Entitäten (einzelne, oder mehrere Wörter) innerhalb des Textes als Informationen enthaltendes Fragment zu identifizieren. Die Entitäten werden dann entsprechenden Terminologien zugeordnet.
All diese Arbeiten stützen sich auf Annotationseditoren, die anhand eines syntaktischen Regelwerks Zuordnung zwischen Entitäten des Textes und domänenspezifischen Konzepten vornehmen.
MetaMap[3] identifiziert Konzepte des UMLS Metathesaurus in Texten, basierend auf linguistischen Prinzipien.
Der UMLS Metathesaurus[4] gilt als ein mächtiges Werkzeug, um medizinische Texte zu annotieren. Das System ist ein multilingualer Thesaurus, der auf dem UMLS (Unified Medical Language System) basiert. Das UMLS umfasst über zwei Millionen Bezeichnungen für ca. 900.000 Konzepte, mit zwölf Millionen Beziehungen zwischen diesen Konzepten, aus mehr als 60 Kategorien biomedizinischer Vokabularien [2].
Das TEXTMARKER System [7] ist ein regelbasiertes System zur Verarbeitung unstrukturierter Texte. Es dient dazu, einen Experten bei der Wissensextraktion zu unterstützen, indem es vielfältige Möglichkeiten bereitstellt, Wissen zu abstrahieren und zu verarbeiten. TEXTMARKER basiert auf UIMA[5], einem Framework zur Analyse unstrukturierter Daten. Es bietet die Möglichkeit Pipelines zu erzeugen, um Entitäten innerhalb eines Textes, als auch deren Beziehungen zu identifizieren.

[3] https://metamap.nlm.nih.gov/
[4] https://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
[5] https://uima.apache.org/

3 Problemstellung

Für das Training eines jeden Klassifikators wird eine Grundwahrheit benötigt. Somit ist die Umsetzung eines Werkzeugs zur Erzeugung dieser der erste Schritt bei der Entwicklung eines klassifikatorbasierten Systems.
In der Domäne des NLP werden zur Erzeugung der Grundwahrheit häufig Annotationseditoren verwendet, die eine möglichst umfangreiche Verwendung ermöglichen sollen [8].
Abhängig von der geplanten Verwendung können diese allerdings einen funktionellen Overhead besitzen, der eine möglichst normierte Klassifikation durch Experten erschwert. Erfolgt die Annotation anhand allgemein bekannter, beispielsweise syntaktischer, Regeln, so erfährt dieser funktionelle Overhead nicht weiter Gewichtung. Sind die Regeln der Annotation allerdings sehr spezifisch, so kann die Annotation durch Experten erst nach einem Training erfolgen. Ziel allerdings ist, mit einem möglichst geringen Trainingsaufwand der Experten, möglichst eindeutige Features zu erhalten.

3.1 Charakteristik des Corpus

Der Corpus bestehen aus insgesamt über 300.000 realen, anonymisierten, patientenspezifischen Befunden und Diagnosen. Diese wurden aus dem Klinikinformationssystem der Augenklinik des Universitätsklinikums Tübingen exportiert.
Ein Element des Corpus besteht aus patienten- und augenspezifischen Befunden zu den anatomischen Regionen eines Auges (Macula, VAA) und den zugehörigen Diagnosen. Die Diagnosen sind sowohl befundspezifisch, als auch, als Teil der Anamnese, über andere Informationssysteme zusammengetragen.
Charakteristisch für die Befunde ist, dass nicht nach einer einheitlichen Vorlage erstellt werden, sondern lediglich als Freitext

in deutsche Sprache mit diversen Fachbegriffen und Abkürzungen vorliegen.
Urheber der Texte sind verschiedene Ärzte der Klinik. Demnach unterscheiden sich die Befunde individuell, sind sprachlich nicht normiert und enthalten unterschiedlichste Schreibstiele und Abkürzungen für ein und den selben Term (Glaskörper: Gk, Glask.).

> Pigmentflecken mit Atrophiebereichen im gesamten HP auch über die GFB nach peripher hinausgehend. Keine Knochenkörperchen. Keine RP.
> Wie RA
> reizfrei, Cat inc

Abbildung 1: Auszug eines Befundes aus dem Corpus.

Da die Befunde im Arbeitsalltag der Mediziner häufig unter Zeitdruck entstehen, beschränken sie sich inhaltlich auf eine exakte Beschreibung der physiologischen Gegebenheiten, sind aber syntaktisch nicht immer korrekt und verzeichnen vereinzelt Tippfehler. Semantisch werden häufig Zeilenumbrüche für die einzelnen Befundungsmerkmale verwendet (Abbildung 1).

4 Analyse vorhandener Lösungen

Ausgehend von der konkreten Problemstellung wurden bei der Konzeption des Systems vorhandene Lösungen untersucht und deren Verwendung mit dem gegebenen Corpus eruiert.
MetaMap stellt ein mächtiges Werkzeug für die automatische Wissensextraktion dar. Allerdings ist es nicht für deutsche Datensätze konzipiert und arbeitete bei mehreren Stichproben auf den gegebenen Datensätzen nicht zuverlässig. Weiter führten die unterschiedlichen Abkürzungen zu falschen Zuordnungen, was einen erhöhten Aufwand durch notwendige Korrekturen bedeuten würde.
Alternativ besteht die Möglichkeit zur manuellen Annotation des Trainingsdatensatzes. Häufig wird hierzu das Konzept des Part Of Speech Taggings (POS-Tagging) verwendet. Es dient der semantischen Analyse der natürlichen Sprache, bei der jedes Wort eines Satzes der entsprechenden morphologischen Kategorie zugeordnet wird. Nach [3] sind die morphologische Dekomposition eines Satzkonstrukts und die Identifikation der einzelnen Bestandteile grundlegende Voraussetzungen für eine erfolgreiche semantische Analyse. Nur wenn diese beiden Vorbedingungen erfüllt sind, kann versucht werden, einen kompletten Ausdruck zu deuten.
Aufgrund der Charakteristika des Corpus wurde von einer automatischen Annotation der Befunde Abstand genommen. Auch das manuelle Tagging mit herkömmlichen POS-Werkzeugen könnte sich, wie auch in [10] erläutert, aufgrund der genannten Eigenschaften des Corpus als problematisch darstellen.

5 Methode

Um den Aufwand dieser Trainingsphase so weit als möglich zu minimieren, wurde ein Annotationseditor entwickelt, der auf die spezifische Problemstellung zugeschnitten ist und die Experten durch Einschränkungen zu einer möglichst normierten Annotationsform bewegt.
Der Annotationseditor soll einen ersten Ansatz für die Projektdurchführung bieten und dient als unabdingbares Werkzeug für die Umsetzung des Gesamtsystems.
Das Gesamtsystem folgt einer Client-Server Architektur. Der Client in Form des Editors bietet die Schnittstelle zur Vorverarbeitung der Befunde. Der Server verfügt über die gesamte Logik für die Bereitstellung der zu annotierenden Befunde, Verwaltung der Selben, Extraktion der Merkmale und Klassifikation dieser.
Letztendlich umfasst die Methode nun, angelehnt an das Vorgehen von Fette et al. [5], die Konzepte als Abstraktionen der textuellen Entitäten in einem Katalog abzubilden (Abbildung 2). Der verwendete Katalog dient als

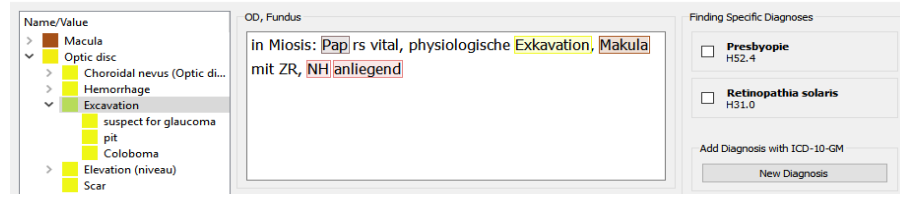

Abbildung 2: Screenshot des Annotationseditors. Links befindet sich der Merkmalskatalog, in der Mitte der zu annotierende Abschnitt des Befundes und rechts die Befundspezifischen Diagnosen.

Referenz für das Tagging der Wörter und kodiert das Domänenwissen. Der Katalog ist als Baum aufgebaut. Die einzelnen Konzepte werden zu den Blättern hin immer feingranularer aufschlüsselt. Der Katalog kann von dem annotierenden Experten erweitert werden. Somit ist die Möglichkeit gegeben, alle in der Realität vorkommenden Konzepte zu fassen. Durch den Editor können in einem Textfenster einzelne Tokens oder Textabschnitte markiert werden und durch Zuweisung zu einem Konzept eine Annotation erstellt werden (Abbildung 2). Eine Annotation besteht demnach aus einer oder mehreren Entitäten, wobei jede Entität die Anker- und Cursorpositionen innerhalb des Textdokuments kodiert.
Durch den Editor können eine oder mehrere Entitäten einem Konzept zugewiesen werden. Zuweisungen von verschiedenen Entitäten zu dem selben Konzept sind zulässig, da ein Konzept mehrmals in einem Text vorkommen kann. Allerdings darf die identische Entität nicht mehreren Konzepten zuweisbar sein. Hierdurch soll die Eindeutigkeit eines Konzepts gewahrt werden.
Um eine möglichst frühzeitige Verifikation der Merkmalsklassifikation zu erreichen werden, angelehnt an [5], bereits gelernte Konzepte mit ihren textuellen Merkmalen in zu annotierenden Befunden automatisch vorannotiert. Der Experte hat somit die Möglichkeit falsche Konzeptzuordnungen zu korrigieren.
Wie bereits eingangs erläutert, umfasst ein Element des Corpus neben den befundspezifischen Diagnosen auch solche, die im Rahmen der Anamnese erfasst wurden. Da diese Diagnosen nicht aus den Befundungstexten hervor gehen, dürfen sie für das Training des Klassifikators nicht verwendet werden. Der Editor verfügt daher über die Funktion, nicht zu berücksichtigende Diagnosen manuell aus dem Corpuselement zu entfernen. Weiter muss berücksichtigt werden, dass sich Diagnosen anhand der textuellen Merkmale eines Befundes ableiten lassen, die im Konkreten Fall nicht gestellt und somit auch nicht vermerkt wurden. Daher können neue Diagnosen anhand des ICD-10-GM Kataloges hinzugefügt werden (Abbildung 2).

6 Bisherige Ergebnisse

Nachdem sich das System aktuell noch in der Entwicklung befindet, konnte mit dem Annotationseditor, aufgrund der verteilten Systemarchitektur, noch keine ausgiebige Evaluation durchgeführt werden.
Fette et al. geben, auf Basis einer Fallstudie, die zeitliche Dauer der Annotation eines kompletten Corpus mit maximal 45, bis minimal 30 Minuten an [5].
Aufgrund der geringen Komplexität des Annotationsverfahrens und der Kürze der einzelnen Befunde, kann davon ausgegangen werden, dass die Annotation eines einzelnen Befundes schneller durchgeführt werden kann.
Bei der ersten Vorstellung des Werkzeuges konnte ein ungeübter Arzt in wenigen Minuten einen kompletten Befund annotieren.
Sobald die Backend-Funktionalität in Gän-

ze vorhanden ist, werden hierzu ausgiebige Tests erfolgen.
Allgemein erregte das System großes Interesse, da die Anwendungsmöglichkeiten über eine reine Diagnoseempfehlung hinaus gehen. Da die Berechnung der ärztlichen Leistung nicht nur anhand der gestellten Diagnosen, sondern auch anhand der, zur Stellung der Diagnosen durchgeführten Methoden erfolgt, ist seitens der Ärzte die Abbildung der methodischen Beschreibungen auf Konzepte wünschenswert.

7 Fazit und Ausblick

Der Editor bietet eine rudimentäre, aber vorerst ausreichende Möglichkeit unkompliziert durch Annotation von Befunden des Corpus eine Grundwahrheit zu schaffen. Diese wird benötigt, um die darunterliegende Kernfunktionalität des Systems zu optimieren und zu evaluieren. Hinsichtlich der Menge der vorhandenen Daten bietet sich zur Optimierung des Workflows in einem späteren Schritt die Überarbeitung des Editors nach Gesichtspunkten der Usability an.
Sobald das System umfassend getestet und evaluiert wurde, soll ein Konzept erarbeitet werden, um es als externes Modul an das KIS der Klinik anzubinden, um direkt bei der Erstellung der Befunde, anhand der dem Text inhärenten Informationen, entsprechende ICD-10 Codierungen bereitzustellen.
Weiter ist der Einsatz als klinikinterne Suchmaschine nach Befunden mit spezifischen Charakteristika möglich.

Literatur

[1] W. Black. *Text Mining*. 2006.

[2] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D–270, 2004.

[3] Y. C. Y. Chen and Z. W. Z. Wang. A semantic method for coding of ICD diagnoses. *3rd International Conference on Biomedical Engineering and Informatics (BMEI)*, (May 1990):2876–2879, 2010.

[4] M. T. Chiaravalloti, R. Guarasci, V. Lagani, E. Pasceri, and R. Trunfio. A Coding Support System for the ICD-9-CM Standard. *2014 IEEE International Conference on Healthcare Informatics*, pages 71–78, 2014.

[5] G. Fette, P. Kluegl, M. Ertl, S. Störk, and F. Puppe. Information Extraction from Echocardiography Records. In *Workshop Notes of the LWA 2011 - Learning, Knowledge, Adaptation*, 2011.

[6] C. Friedman, T. C. Rindflesch, and M. Corn. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of biomedical informatics*, 46(5):765–73, oct 2013.

[7] P. Kluegl, M. Atzmueller, and F. Puppe. Textmarker: A tool for rule-based information extraction. *Proceedings of the Biennial GSCL Conference 2009, 2nd UIMA@GSCL Workshop*, pages 233–240, 2009.

[8] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, F. Puppe, P.-d. B. Georg, and F. Frank. UIMA Ruta Workbench: Rule-based Text Annotation. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 29–33, 2014.

[9] J. Medori and C. Fairon. Machine learning and features selection for semi-automatic ICD-9-CM encoding. In *Second Louhi Workshop on Text and Data Mining of Health Documents (Louhi-10)*, number June, page 84, 2010.

[10] A. Sotelsek-Margalef and J. Villena-Román. MIDAS: An Information-Extraction Approach to Medical Text Classification. *Procesamiento del lenguaje Natural*, 41:97–104, 2008.

Weiterentwicklung eines Fingertrackingsystems zur Steuerung holografischer Benutzeroberflächen

Alexander Niedel
Reutlingen University
Alexander.Niedel@Student.Reutlingen-University.DE

Abstract

In Situationen in denen keine Anzeigegeräte wie Monitore oder Touchscreens vorhanden sind könnten grafische Benutzeroberflächen als Hologramm, frei schwebend, im Raum dargestellt werden. Die Interaktion mit dieser Art von Benutzeroberflächen gestaltet sich allerdings schwierig da keine tatsächliche Berührung wie bei Touchscreens stattfindet. Eine Möglichkeit die Interaktion trotzdem zu realisieren bietet die Verwendung von Fingertrackingsystemen. In dieser Ausarbeitung wird der Prototyp eines holografischen Projektionssystems vorgestellt und die Weiterentwicklung des Fingertrackings beschrieben, die die Interaktion robuster und zuverlässiger gestalten soll.

Schlüsselwörter

Fingertracking, Hologramm, Tiefenkamera, Interaktion, grafische Benutzeroberfläche

CR-Kategorien

Input devices, Virtual device interfaces, Virtual reality, Edge and feature detection, Projections, Depth cues, Tracking

Betreuer Hochschule: Prof. Dr. rer. nat. Uwe Kloos
Hochschule Reutlingen
Uwe.Kloos@Reutlingen-University.de
Betreuer Firma: Dr. Matthias Bues
Fraunhofer IAO
matthias.bues@iao.fraunhofer.de

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

1 Einleitung

Computerinhalte, angezeigt auf Hologrammen, frei schwebend im Raum dargestellt anstatt auf fest verbauten Monitoren. Das sind gern genutzte Elemente in futuristisch dargestellten Filmen. Benutzeroberflächen, Bilder und Videos werden direkt in den Raum projiziert. Interagiert wird meist durch Berührung der Hologramme mit den Fingern. Das Konzept der Interaktion ist dem von Touchdisplays nachempfunden.

Die Verwendung von Hologrammen würde einige Vorteile mit sich bringen. Da keine sichtbare Hardware verwendet werden muss können Benutzeroberflächen nur bei Bedarf angezeigt werden und belegen keinen festen Platz im Umfeld des Benutzers. Auch kann die Position der Visualisierung, im Rahmen der Möglichkeiten des Projektionssystems, frei gewählt und verändert werden. Es findet bei der Interaktion keine tatsächliche Berührung statt. Es kommt daher auch nicht zur Verschmutzung der Hardware durch die Finger wie es bei Touchdisplays der Fall ist. Beispielsweise im medizinischen Bereich dürfte dieser Vorteil interessant sein. Durch die fehlende Berührung fällt aber auch das haptische Feedback, das schon

Touchdisplays teilweise nur noch simulieren, komplett weg. Hierfür müssen alternative Konzepte für die Rückmeldung an den Benutzer gefunden werden.

Was in unserer Vorstellung der Zukunft bereits einen festen Platz in unserem Alltag zu haben scheint kann heutzutage schon teilweise realisiert werden.

2 Holografische Darstellung von Benutzeroberflächen

Um eine Benutzeroberfläche holografisch darzustellen wird ein Projektionssystem verwendet, welches das Bild eines LCD-Displays über einen gebogenen Spiegel perspektivisch so verzerrt, das es vom Betrachter als im Raum schwebend wahrgenommen wird. Dieser optische Effekt wird auch stereoskopische Anamorphose [1] genannt und eignet sich für diesen Anwendungsfall da das Bild je nach Position des Betrachters von dem Spiegel immer perspektivisch korrekt verzerrt wird.

Abbildung 1: Prototyp HoloBeamer

Der verwendete Prototyp für diese holografische Projektion wurde von der Firma 3D-Around [2] hergestellt. In einem Metallgehäuse sind aktiv gekühlte LEDs als Lichtquelle für ein LCD-Display mit einer Auflösung von 1024x768 verbaut. Das Display befindet sich im vorderen Teil des Gehäuses und wird vom Benutzer nicht direkt gesehen. Von oben blickt der Benutzer durch eine verspiegelte Glasplatte in das Innere des Gehäuses auf dem an der Rückwand verbauten, gebogenen Spiegel. Der Spiegel reflektiert das Bild des Displays und verzerrt es so, dass beim Benutzer der Eindruck entsteht als würde es auf vorderen Rand des Gehäuses schweben.

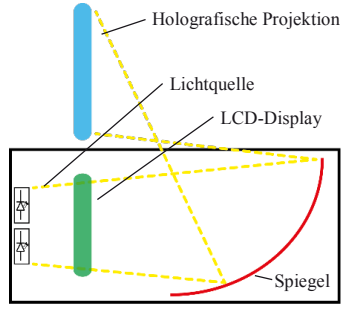

Abbildung 2: Funktionsweise der holografischen Projektion

Auf Abbildung 1 ist der Effekt visualisiert. Obwohl keine stereoskopische Darstellung möglich ist, lässt sich das Ergebnis erahnen.

Das Projektionssystem kann wie ein Monitor an einem PC angeschlossen und betrieben werden. Es unterliegt allerdings der Einschränkung, dass sich der Benutzer so platzieren muss, dass er die ganze Projektion des Displays in seinem Sichtfeld hat. Werden Teile der verzerrten Spiegelung verdeckt oder abgeschnitten, funktioniert die holografische Wahrnehmung des Displays nicht mehr. Für das Anwendungsszenario dieses Prototyps bedeutet dies, dass das Projektionssystem in Tischhöhe angebracht werden muss. Der Benutzer steht davor und blickt darauf hinab. Der Interaktionsbereich der sich ca. 2 - 20 cm über dem Tiefensensor befindet ist dabei in direkter Reichweite der Hand des Benutzers.

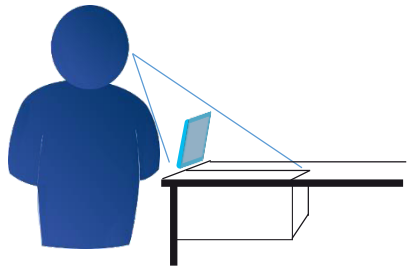

Abbildung 3: Anwendungsszenario

Die Interaktion mit der holografischen Projektion wird in der Demonstrationsanwendung des Prototyps über eine LeapMotion [3] realisiert. Eine Fingertrackingsoftware errechnet eine virtuelle Ebene im Koordinatensystem der LeapMotion. Diese virtuelle Ebene entspricht der Position an der die holografische Projektion vom Benutzer wahrgenommen wird. Die Trackingdaten werden dann auf Schnittpunkte mit dieser virtuellen Ebene untersucht. So werden Berührungen des Benutzers mit der holografischen Projektion erkannt und können für die Interaktion verwendet werden. Darüber hinaus können Effekte wie das visuelle Hervorheben eines Elements beim Annähern eines Fingers umgesetzt werden da im Gegensatz zu Touchdisplays die Fingerposition bereits im nahen Bereich der Benutzeroberfläche bestimmt werden kann. Die Demonstrationsanwendung besteht aus einer kleinen Bildergalerie. Der Benutzer kann sich beim Berühren des rechten oder linken Rands der holografischen Projektion das nächste oder vorherige Bild anzeigen lassen.

3 Eignung des Prototyps zur Steuerung von Lichtanlagen

Das Projekt ist Teil des BMBF-Verbundprojekts OLIVE [4] und soll dabei helfen Möglichkeiten zur Steuerung der gesamten Beleuchtungssituation in einem Raum zu finden.

In ersten Versuchen sollte die Eignung eines solchen holografischen Projektionssystems zur Steuerung von simplen Beleuchtungsanlagen getestet werden. Dazu wurde eine Benutzeroberfläche gestaltet auf der vier Buttons angewählt und deren Werte mit einem Schieberegler verändert werden können. Die sehr einfach gehaltene Benutzeroberfläche soll simple Bedienelemente wie Lichtschalter ersetzen und erweiterte Interaktionsmöglichkeiten bieten. Das Interaktionskonzept sieht zunächst eine einfache Steuerung mit Touchdisplay-ähnlichem Verhalten vor. Später soll das Interaktionskonzept durch Gesten und andere, möglicherweise neue, Methoden erweitert werden.

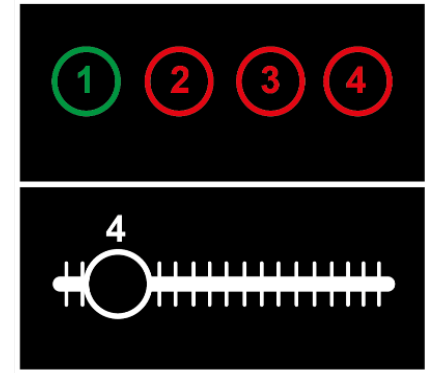

Abbildung 4: Screenshot der Benutzeroberfläche

Um den Benutzer bei der Interaktion nicht zu behindern müssen die Funktionen vereinfacht oder mit gleichem Aufwand genutzt werden können.

Die umzusetzenden Interaktionen wie das Bedienen von (Licht)-Schaltern oder Schiebereglern müssen aufgrund des gewohnten Gebrauchs im Alltag auf Anhieb funktionieren. Die Interaktion soll nicht erst gelernt werden müssen sondern möglichst intuitiv vom Benutzer ausgeführt werden. Als Wissensgrundlage des Benutzers kann von dem gelernten Umgang mit Touchdisplays und realen Schaltern sowie grafischen Benutzeroberflächen ausgegangen werden.

Die LeapMotion erwies sich zum Erfassen der Fingerpositionen als ungeeignet. Vor allem im nahen Bereich traten zu starke Abweichungen der Tiefenwerte des Sensors auf. Dadurch wurden die Berührungen mit der holografischen Darstellung nicht zuverlässig genug erkannt.

Da das geplante Interaktionskonzept aber eine zuverlässige Erfassung der Benutzereingabe voraussetzt muss das Fingertracking neu umgesetzt und robuster gestaltet werden.

4 Weiterentwicklung des Fingertrackingsystems

Um das Fingertrackingsystem robuster zu gestalten werden Soft- und Hardware ausgetauscht. Eine genauere Erfassung der Tiefenwerte im nahen Kamerabereich und ein zuverlässigeres Tracking der Fingerspitze des Benutzers sind Ziel dieser Änderungen. Die Methode zur Erkennung von Berührungen mit der virtuellen Darstellungsebene der holografischen Projektion kann weitgehend übernommen werden.

4.1 Hardware

Als Tiefensensor kommt die Creative Senz3D [5] zum Einsatz. Sie liefert ein Farbbild mit einer Auflösung von 640x480 Pixeln und ein Tiefenbild mit immerhin 320x240 Pixeln. Sie nutzt die Time-Of-Flight Technologie [6] zum Erfassen der Tiefenwerte. Dabei wird ein Lichtimpuls im Infrarot Bereich ausgesendet und die Zeit für jeden Pixel im Tiefenbild der Kamera berechnet die das Licht benötigt um reflektiert zu werden. Aus den gewonnenen Daten wird die Distanz zur Kamera bestimmt. Die Bilder und Tiefenwerte können, anders als bei der LeapMotion, roh aus der Kamera ausgelesen und selbst verarbeitet werden. Dies ermöglicht eine eigene Implementation der Fingerspitzenerkennung.

Die Kamera soll wie die davor benutze LeapMotion am vorderen Rand des Gehäuses befestigt werden und die Finger des Benutzers von unten erfassen. Auf die Art kann sowohl das Projektionssystem als auch das Trackingsystem unter einer Glasoberfläche für den Benutzer unsichtbar verbaut werden.

Aufgrund der Größe der Tiefenkamera sind Änderungen am Gehäuse des Prototyps geplant um sie tiefer platzieren zu können. Der Momentane Abstand des Sensors zum unteren Rand des Interaktionsbereichs von ca. 2 cm liegt gerade noch in einem Bereich in dem die Kantenerkennung überhaupt möglich ist.

4.2 Software

Geplant ist eine Software die Fingerspitzen erkennt ohne dabei ein Handmodell zu verwenden. Das ist notwendig da die Interaktionen in einem sehr nahen Bereich der Kamera stattfinden und sich die Hand des Benutzers oft nicht im Bild der Kamera befindet. Die gewonnenen Positionsdaten sollen analysiert und so Berührungen mit der holografischen Projektion erkannt werden. Die Interaktionen des Benutzers sollen an die Benutzeroberfläche weitergegeben und dort visualisiert und verarbeitet werden.

Die Software gliedert sich also in einen Trackingteil, der Positionsdaten generiert und analysiert, und in einen Teil in dem die Visualisierung der Benutzeroberfläche und die Verarbeitung der Benutzereingaben umgesetzt werden. Durch diese Aufteilung können beliebige Benutzeroberflächen mit derselben Trackingsoftware gesteuert werden.

Die Kommunikation der Softwareteile untereinander erfordert die Implementierung einer geeigneten Schnittstelle und den Entwurf eines passenden Protokolls.

4.2.1 Trackingteil

Die Trackingsoftware nutzt das Intel Perceptual Computing SDK [7] zum Erfassen der Farb- und Tiefenbilder und OpenCV [8] für die Bildverarbeitung und die Fingererkennung. Die in dem SDK von Intel mitgelieferte Fingererkennung basiert auf einem Handmodell und erfordert die Positionierung der Kamera vor dem Benutzer. Da das im Anwendungsfall nicht möglich ist kann sie nicht verwendet werden.

Eine eigene Methode zur Fingererkennung aus den rohen Sensordaten wird benötigt.

Um brauchbare Bilder für eine eigene Fingererkennung zu erhalten werden zunächst die Farb- und Tiefenwerte roh ausgelesen. Bei dem Farbbild wird der Hintergrund mittels eines Tiefenfilters entfernt um Störungen durch Objekte außerhalb des Interaktionsbereichs möglichst auszublenden. Dafür werden jedem Pixel des Farbbilds die entsprechenden Tiefenwerte zugeordnet.

Die Auflösung der erfassten Tiefenwerte beträgt nur 320x240 Pixel und kann nicht eins zu eins auf die Pixel des Farbbilds mit der Auflösung von 640x480 übertragen werden. Um die Tiefenwerte trotzdem auf die Pixel des Farbbilds zu übertragen muss entweder die Auflösung des Farbbilds verringert oder jeder Tiefenwert auf mehreren Farbpixeln abgebildet werden. Um die Bildqualität für das geplante Fingertracking möglichst hoch zu halten wird die Auflösung des Farbbilds nicht reduziert. Entstehende Lücken in den Tiefeninformationen werden mit Tiefenwerten der Nachbar-Pixel aufgefüllt. Fehler die durch dieses Mapping einzelner Tiefenwerte auf mehrere Pixel entstehen befinden sich nur im direkten Umfeld der zu trackenden Objekte. Dadurch sollte es nur zu minimalen Abweichungen kommen die nicht weiter ins Gewicht fallen. Pixel mit zu großem Abstand zur Kamera können nach dem Mapping der Tiefenwerte schwarz eingefärbt werden. Somit werden nur Objekte die sich im Interaktionsbereich befinden im nächsten Schritt berücksichtigt.

Mit der Canny-Kantenerkennung [9] aus dem OpenCV Framework werden die Konturen des Bildes herausgefiltert und isoliert dargestellt.

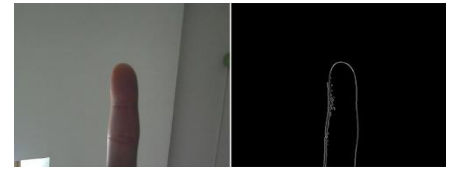

Abbildung 5: Original Bild der Kamera und Ergebnis der Kantenerkennung

Zusammenhängende Konturen werden mittels der Methode findContours(…) separat abgespeichert. Dies ermöglicht ein simples Tracking der vordersten Fingerspitze indem der Punkt der Fingerkontur mit dem größten y-Wert ausgelesen wird. Die Distanz zur Kamera dieses Punktes wird aus den Tiefeninformationen separat ermittelt. Mit diesen Daten kann die Position des Punkts im Koordinatensystem der Kamera bestimmt werden.

Über einen Kalibrierungsvorgang, bei dem der Benutzer 3 Punkten auf der holografischen Projektion Positionswerte im Koordinatensystem des Fingertrackings zuordnet indem er sie mit der Fingerspitze berührt, wird die virtuelle Ebene errechnet auf der das Hologramm wahrgenommen wird. Diese virtuelle Ebene befindet sich ebenfalls im Koordinatensystem des Fingertrackings. Eine Projektion der Trackingdaten in ein anderes Koordinatensystem ist nicht notwendig. Diese Kalibrierung muss einmalig von einem Benutzer durchgeführt werden. Die Daten zur Berechnung der virtuellen Ebene werden in einer Datei abgelegt und gespeichert. Trotzdem ist eine Wiederholung der Kalibrierung bei jedem Benutzerwechsel sinnvoll, da die Position der Projektionsebene von der Wahrnehmung des Benutzers abhängt. Berührungen der Fingerspitze mit der virtuellen Projektionsebene können dann erkannt und als Touchevent oder Klick behandelt werden. Außerdem ist es möglich Einen Cursor für eine sich nähernde Fingerspitze auf der Bedienoberfläche darzustellen oder Mouseover-Effekte zu realisieren. Anders als bei Touchinterfaces ist die Position des Fingers schon vor dem Klick bekannt. Das Konzept zur Erkennung von Berührungen mithilfe der virtuellen Ebene wurde aus der mitgelieferten Demonstrationssoftware des Prototyps weitgehend übernommen.

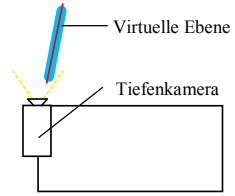

Abbildung 6: Tiefenkamera und virtuelle Ebene

4.2.2 Benutzeroberfläche und Kommunikation

Bei der Benutzeroberfläche handelt es sich um einen Prototyp der mit Web-Technologie umgesetzt wurde. HTML, CSS und JavaScript Code wurden für die Implementation verwendet. Das Layout und die Elemente sowie ihr Verhalten bei der Interaktion können schnell und mit wenig Aufwand erstellt werden. Trotzdem bieten sie eine Umfangreiche Funktionalität. Die Verwendung von Web-Technologien ermöglicht außerdem den Einsatz in verteilten Systemen, da die Darstellung der Benutzeroberfläche vom Betriebssystem unabhängig in einem Internetbrowser stattfindet. Trackingsoftware und Benutzeroberfläche müssen also nicht zwingend auf demselben Gerät laufen. Für die Kommunikation mit der Trackingsoftware ist eine einfache Anbindung an Netzwerk oder Internet möglich.

Um eine bidirektionale Verbindung zwischen Trackingsoftware und Benutzeroberfläche zu schaffen und so Push-Nachrichten senden zu können wird ein WebSocket [10] Server implementiert. Dieser leitet alle Nachrichten die er erhält an jeden verbundenen Client weiter. Der Server ist in C-Code geschrieben und läuft als Stand-Alone Anwendung. Sowohl die Trackingsoftware als auch die Benutzeroberfläche melden sich über einen WebSocket Client bei dem Broadcast-Server an und können rohe Textnachrichten austauschen. Dieser Client kann in der jeweiligen Programmiersprache implementiert werden.

Die Trackingsoftware sendet Mouseover und Klickevents mit entsprechenden Positionsdaten an die Benutzeroberfläche. Dabei werden x-, y- und z-Werte angepasst um der Zielauflösung des Projektionssystems von 1024x768 Pixeln zu entsprechen. An der den Positionsdaten entsprechenden Stelle auf der Benutzeroberfläche werden die Mausklicks und Mouseoverevents mittels JavaScript Funktionen nur noch simuliert. Auf diese Weise muss bei der Implementation der HTML-Benutzeroberfläche keine weitere Rücksicht auf die spezielle Interaktion genommen werden. Die Funktionalität zum Einschalten des Lichts oder andere Aktionen die durch die Bedienelemente der Benutzeroberfläche angeboten werden, werden von ihr selbst implementiert. Mit der Integration eines WebSocket Clients und der Funktion zur Simulation der Trackingevents kann so jede beliebige HTML-Seite als potentielle holografische Benutzeroberfläche nutzbar gemacht werden.

Erfolgreich angeklickte Elemente werden mit entsprechenden Statusinformationen der Trackingsoftware zurückgemeldet.

5 Erkenntnisse und Weiterführende Arbeiten

Das weiterentwickelte Fingertrackingsystem ist in der Lage Interaktionen eines Benutzers mit einer holografischen Projektion zuverlässiger zu erkennen. Dies ist zu einem großen Teil auf der Fingererkennung ohne die Nutzung eines Handmodells zurückzuführen.

Um genauere Aussagen über die Benutzbarkeit und Akzeptanz dieses Konzepts zur Interaktion mit holografischen Projektionen machen zu können müssen zunächst aussagekräftige Nutzertests durchgeführt und ausgewertet werden. Das Konzept zur Interaktion mit Touchdisplay-ähnlichem Verhalten holografischer Benutzeroberflächen lässt sich aber Umsetzen. Weitere Verbesserungen an der Software, wie zum Beispiel eine verbesserte Aufbereitung der Daten vor der Kantenerkennung, müssen vorgenommen werden um die gewünschte Robustheit bei der Interaktion zu erreichen.

Das System bietet noch viel Raum für Erweiterungen. Mit der Integration einer Gestenerkennung könnte die Funktionalität erweitert und damit mehr Möglichkeiten für neue Interaktionskonzepte geschaffen werden. Neue verbesserte Sensoren zur Tiefenerkennung würden die Präzision der Trackingdaten noch erhöhen was sich bei der Benutzung spürbar positiv auswirken könnte.

Auch wenn die Technologie zur Projektion von Hologrammen noch am Anfang steht konnte gezeigt werden dass die Interaktion mit Ihnen schon heute durchaus realisierbar ist.

6 Literaturverzeichnis

[1] Stereoskopische Anamorphosen, Website http://www.3dwebsite.de/de/html/anamorphosen.html, Accessed 11 November 2015.

[2] 3D Around GmbH, Website http://www.3daround.com, Accessed 11 November 2015.

[3] Leap Motion Controller, Website http://store-eur.leapmotion.com/products/leap-motion-controller, Accessed 11 November 2015.

[4] Optimierte Lichtsysteme zur Verbesserung von Leistungsfähigkeit und Gesundheit (OLIVE), BMBF-Verbundprojekt (FKZ: 13N13158) Projektsteckbrief: http://www.photonikforschung.de/fileadmin/Verbundsteckbriefe/4._LED/Projektsteckbrief_OLIVE.pdf, Accessed 11 November 2015.

[5] Creative Senz3D: Technical Specifications, Website http://support.creative.com/kb/ShowArticle.aspx?sid=120808, Accessed 11 November 2015.

[6] Miles Hansard, Seungkyu Lee, Ouk Choi, Radu Horaud. Time of Flight Cameras: Principles, Methods, and Applications. Springer, pp.95, 2012, SpringerBriefs in Computer Science, ISBN 978-1-4471-4658-2.

[7] Intel Perceptual Computing SDK-Dokumentation, Website https://software.intel.com/sites/landingpage/perceptual_computing/documentation/html/, Accessed 11 November 2015.

[8] OpenCV Framework, Website http://opencv.org/, Accessed 11 November 2015.

[9] J. Canny. A Computational Approach to Edge Detection, IEEE Trans. on Pattern Analysis and Machine Intelligence, 8(6), pp. 679-698 (1986).

[10] RFC – Standard: The WebSocket Protocol, Website http://datatracker.ietf.org/doc/rfc6455/?include_text=1, Accessed 11 November 2015.

Konzept für ein portables System zur Müdigkeitserkennung mit Körpersensoren *

Paul Pasler
Reutlingen University
Paul.Pasler@Student.Reutlingen-University.DE

Abstract

Mit Fortschreiten der Technik, verbreiten sich Fahrerassistenzsystemen immer weiter. Besonders der Teilbereich der Müdigkeitserkennung hilft schwere Unfälle zu vermeiden. Die Müdigkeitserkennung mit Body-Sensorik liefert sehr gute Ergebnisse, scheitert aber in der Praxis häufig auf Grund seines invasiven Charakters. Für die vorgelegte Arbeit werden Forschungsergebnisse aus diesem Bereich evaluiert und daraus im Simulationsumfeld der Reutlingen University ein Konzept entwickelt, dass Körperfunktionen überwacht und diese auswertet, ohne den Fahrer zu beeinträchtigen. Weiterhin wird die Möglichkeit einer einfachen Portierung der Anwendung vom Simulator in ein echtes Fahrzeug geprüft. Das vorgestellte Konzept, soll somit ein Höchstmaß an Genauigkeit, Tragekomfort und Mobilität vereinen.

Schlüsselwörter

Advanced Driver Assistance System (ADAS), Fahrerassistenzsystem, Müdigkeitserkennung

CR-Kategorien

A.0 [**ACM**]: Experimentation

*

Betreuer Hochschule: Prof. Dr. Martinez
Hochschule Reutlingen
Natividad.Martinez@Reutlingen-University.de
Wissenschaftliche Vertiefungskonferenz 2015
Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

1 Einleitung

Fristeten Fahrerassistenzsysteme vor wenigen Jahren ein Nischendasein in Oberklassewagen, werden sie immer günstiger und beliebter. Mittlerweile halten sie auch in Mittelklasse- und Kleinwagen Einzug. Die Unternehmensberatung Strategy Analytics geht in den nächsten Jahren von einer Versechsfachung von verbauten Fahrerassistenzsysteme aus [1]. Denn sie erhöhen den Komfort und helfen bei der Vermeidung schwerer Unfälle. So schätzt die Boston Consulting Group, dass durch den flächendeckenden Einsatz von Fahrerassistenzsysteme, die Unfallrate in den USA um bis zu 28% zurückgehen könnte [2]. Müdigkeitserkennung ist eines dieser Systeme zur Vermeidung von müdigkeitsbedingter Unachtsamkeit oder Sekundenschlaf. Beispielsweise rät die Müdigkeitserkennung „Attention Assist“ von Daimler dem Fahrer, zu gegebenen Anlass, eine Pause einzulegen und zeigt ein Kaffeesymbol im Cockpit an [3]. Denn laut dem Deutschen Verkehrssicherheitsrat zählt Müdigkeit, neben überhöhter Geschwindigkeit, zu den häufigsten Unfallursachen und ist damit für jeden fünften schweren Unfall verantwortlich [4]. In einer Studie der amerikanischen „National Sleep Foundation“ [5], gab die Hälfte der Befragten an, dass sie schon einmal schläfrig gefahren seien und fast jeder dritte sogar kurz am Lenkrad eingeschlafen war. Dies zeigt die Wichtigkeit einer Erkennung von Müdigkeit und einer Meldung an den Fahrer.

Abbildung 1: Skizze des Systemaufbaus: Körpersensoren (Elektroenzephalografie / Elektrokardiogramm) liefert Daten an die Applikation und ein Feedback-Device warnt den müden Fahrer. Bild zeigt den Fahrsimulator der Reutlingen University.

Um das Risiko eines Unfalls auf Grund von Übermüdung zu senken, soll langfristig ein multimodales System zu Müdigkeitserkennung entwickelt werden (Siehe Abb. 1). Solche Systeme existieren bereits, es fehlt jedoch oftmals an Komfort und Portabilität. Ziel dieser Arbeit ist es, aktuelle Arbeiten zu diesem Thema zu sichten und ein Konzept für ein solches System mit Körpersensoren zu erstellen. Die Körpersensoren messen Signale direkt am Körper und können somit sofort auf Veränderungen reagieren. Ein Algorithmus versucht an Hand der Messdaten zu erkennen, ob der Fahrer übermüdet. Diese Systeme müssen richtig und genau funktionieren, sodass die Sicherheit zu jeder Zeit gewährleistet ist. Da die Sensoren direkt am Körper anliegen, können sie den Fahrer beeinträchtigen. Das Problem der invasiven Sensoren soll weitestgehend eliminiert und den Fahrer wenig bis gar nicht stören. Feldversuche eignen sich nicht zur Entwicklung eines solchen Systems, da Eigen- und Fremdgefährdung eines übermüdeten Fahrers nicht vertretbar sind. Das System soll darum im Simulationsumfeld der Reutlingen University entwickelt und getestet werden. Dennoch müssen die Ergebnisse einem Test im Straßenverkehr standhalten, da es unter Umständen zu anderen Signalen, beispielsweise aufgrund von erhöhtem Stress, kommen kann. Darum soll das System später leicht in ein echtes Fahrzeug portiert und validiert werden können. Gelingt dies, kann es zudem mit anderen Systemen gekoppelt zu werden, um das Ergebnis insgesamt zu verbessern. Damit hilft das vorgestellte Konzept, den Fahrer vor einer drohenden Müdigkeit zu warnen und so schwere Unfälle zu vermeiden.

Die Ausarbeitung gliedert sich folgendermaßen. Im Kapitel 3 werden verschiedene Forschungsergebnisse zur Müdigkeitserkennung aufgezeigt und in Kapitel 4 verglichen. Das Konzept eines portablen Systems zur Müdigkeitserkennung mit Körpersensoren wird im Kapitel 5 vorgestellt. Der Versuchsaufbau und das Testszenario im Simulationsumfeld der Reutlingen University ist Thema von Kapitel 6. Das Ergebnis und weitere Schritte werden in Kapitel 7 beschrieben. In den anschließenden Absätzen werden Grundlagen für die kommenden Kapitel erläutert.

2 Grundlagen

In den folgenden Abschnitten werden Grundlagen für das Verständnis der weiteren Kapitel gelegt. Es wird ein grober Überblick zu Fahrerassistenzsysteme, Müdigkeitserkennung und Körpersensoren gegeben.

2.1 Fahrerassistenzsysteme

Fahrerassistenzsysteme erhöhen den Komfort bzw. die Sicherheit des Fahrers. So führen Einparkassistent, Geschwindigkeitsregelanlage oder Navigation zu einer deutlich entspannteren Fahrt. Spurhalte-, Spurwechsel- oder Notbremsassistent wiederum unterstützen bei potentiell gefährlichen Manövern. Auch die Müdigkeitserkennung fällt in die zweite Kategorie (mehr dazu in Kapitel 3).

Kompaß [6] unterteilt Fahrerassistenzsysteme, gemessen an der Reaktionszeit, in Planung, Führung und Stabilisierung. Hierbei fällt beispielsweise Navigation in die Planungsebene, da die Berechnung der Route mit unter mehrere Minuten brauchen kann. Auf Führungsebene werden dem Fahrer Empfehlungen und Warnungen innerhalb weniger Sekunden mitgeteilt, auf die er dann reagieren kann. Greift das System selbständig in den Fahrprozess ein, muss dies meist innerhalb von Millisekunden geschehen und dient oftmals zur Stabilisierungen, wie beispielsweise bei einem Fahrdynamik-Regelsystem.

Ein Fahrerassistenzsystem kann auf verschiedenste Arten mit dem Fahrer kommunizieren. Es handelt sich um eine klassische Human-Computer-Schnittstelle. Am gebräuchlichsten, auch für sonstige Warnungen, sind schon seit längerem Optische und Akustische Signale. Aber auch Vibrationen in Lenkrad und Sitz zeigen gute Ergebnisse, wenn zwischen Signal und Nachricht ein Zusammenhang besteht (beispielsweise Vibriert das Lenkrad bei verlassen der Spur). Bertoldi und Filgueiras [7] beschreiben hierzu die verschiedenen Anwendungsgebiete und Unterschiede.

Jeder Automobilhersteller entwickelt mittlerweile seine eigenen Fahrerassistenzsysteme. Datenerhebung (Sensoren), Berechnung und Kommunikation werden vom Fahrzeug selbst durchgeführt. Durch die Abschottung des Fahrzeugs sind Fahrzeugdaten nicht öffentlich zugänglich und können nur schwer von Außenstehenden genutzt werden.

Für wissenschaftliche Arbeiten bleibt entweder eine Kooperation mit Automobilherstellern oder das Ausweichen auf andere Geräte, wie ein Smartphone und die Nutzung von Daten aus dem Internet (beispielsweise Kartendienste). Chen et al. [8] und You et al. [9] verfolgten diesen Ansatz. Smartphone bieten durch ihren hohen Verbreitungsgrad eine günstige Alternative zu eingebauten Systemen, können jedoch nicht auf Daten des Fahrzeugs zugreifen und müssen einfache Daten, wie beispielsweise Geschwindigkeit, selbst berechnen.

2.2 Müdigkeitserkennung

Die Erkennung von Übermüdung kann wiederum auf ganz verschiedene Arten gelöst werden. Ein Ansatz versucht über Körpersignale herauszufinden, ob es Anzeichen für Müdigkeit gibt. Wohingegen mit der Analyse des Fahrverhaltens das Selbe mit Sensoren an und im Auto realisiert wird. Bei der Erkennung über Körpersignale, können wiederum Körpersensoren oder kamerabasierte Computer-Vision (CV) Techniken zur Überwachung des Fahrers genutzt werden (siehe Abb. 2). Zu unterscheiden ist weiterhin die physische und psychische Müdigkeit, welche sich jedoch beide negativ auf die Fähig-

keiten des Fahrers auswirken. Alle Verfahren, die auf Sensoren am Körper, die extra angezogen werden (bspw. ein Pulsmesser am Ohr, Elektroenzephalografie) werden, als invasive Verfahren bezeichnet.

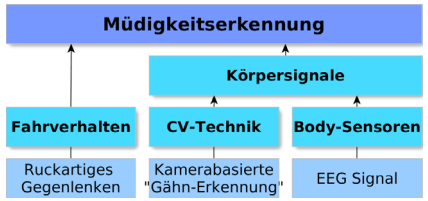

Abbildung 2: Einteilung der Systeme zur Müdigkeitserkennung

Allen System gemein ist die Nutzung von Klassifizierungs- bzw. Machine-Learning-Algorithmen. Die gesammelten Daten geben nur Hinweise und sind kein Garant für eine Erkennung von Müdigkeit. Müdigkeitserkennungssysteme wandeln hier auf einem schmalen Grad, da es zum einen um die Verhinderung schwerer Unfälle geht, zum anderen aber ein falsch auslösendes System die Akzeptanz vermindert und im schlimmsten Fall zu einer Deaktivierung führt. Um falsche Erkennungen weiter zu minimieren, werden oftmals mehrere Ansätze kombiniert.

In der Praxis setzen Automobilhersteller wie Daimler [3] und Volkswagen, sowie Automobilzulieferer wie Bosch [10] auf die Analyse des Fahrverhaltens. Insbesondere Spurhalten und ruckartiges Gegenlenken scheinen ein signifikantes Indiz für beginnende Übermüdung zu sein. Weiterhin sind externe Geräte und einige Apps für Smartphones erhältlich.

2.3 Körpersensoren

Die meisten Körpersensoren messen elektrische Signale eines Körpers, wie den Puls, Temperatur oder Impulse des Gehirns. Meistens werden sie direkt am oder im Körper eingesetzt.

Bei der Elektroenzephalografie (EEG) werden Elektroden auf der Kopfhaut angebracht und damit die Aktivität des Gehirns gemessen. Sie wird in der Medizin für die Diagnose von Epilepsie oder bei Koma-Patienten eingesetzt. Zudem findet sie in Schlaflabors Anwendung, um verschiedene Schlafphasen zu erfassen. Der Zusammenhang von Schlaf und Hirnaktivität kann auch bei der Müdigkeitserkennung in Fahrzeugen genutzt werden, um beispielsweise ein drohenden Sekundenschlaf zu erkennen [11].

Das Elektrokardiogramm (EKG) misst die Herzspannungskurve und stellt die Aktivität des Herzmuskels dar. So lassen sich vielfältige Aussagen über den Zustand des Herzens machen. Weiterhin können Herzrhythmus und -frequenz Hinweise auf eine einfallende Müdigkeit des Fahrers geben. Für die Messung des EKG Signals existieren mehrere Methoden. Im medizinischen Bereich werden Elektroden an verschiedenen Körperstellen geklebt. Weiterhin werden, vor allem im Sport, Sensoren mit Gummibändern an Brust oder Handgelenk befestigt.

Eine weitere Möglichkeit Körperfunktionen aufzuzeichnen ist die Elektrookulografie (EOG). Hierbei kann die Bewegung der Augen bzw. das Ruhepotential der Netzhaut gemessen werden. Dazu werden Elektroden entweder rechts und links oder oben und unter dem Auge angebracht.

Grundlagen von Fahrerassistenzsystem, Müdigkeitserkennung und Körpersensoren waren Thema des vergangenen Kapitels. Für die Realisation von Systemen zur Müdigkeitserkennung existieren verschiedene Arbeiten, diese werden im nächsten Kapitel vorgestellt.

3 Stand der Technik

Müdigkeit senkt die Konzentrationsfähigkeit des Fahrers, kann zu einer erhöhten Reaktionszeit und Fehleinschätzungen führen. Dies stellt beispielsweise der Deutsche Verkehrssicherheitsrat in einem Beschluss von 2009 [12] fest. Ursachen können wenig Schlaf, lange Fahrzeiten, Medikamente oder Alkohol sein. Systeme zur Müdigkeitserkennung versuchen an Hand verschiedener Daten und Sensoren, frühzeitig zu erkennen, ob der Fahrer gerade Anzeichen einer bevorstehenden Müdigkeit zeigt und empfiehlt eine Pause (beispielsweise [3]). Dabei soll nicht nur während eines Micro- oder Sekundenschlafs, sondern schon früher gewarnt werden. Hierfür existieren unterschiedliche Forschungsergebnisse, die im folgenden vorgestellt werden. Eine ausführliche Übersicht findet sich auch bei Sahayadhas et al. [13].

Systeme die das Fahrverhalten analysieren sind in der Praxis weit verbreitet und werden von den meisten Automobilherstellern eingesetzt. Leider existieren kaum öffentlich zugängliche Arbeiten zu diesem Ansatz von Müdigkeitserkennung, da es sich um interne Entwicklungen handelt.

Andere Ansätze beobachten den Fahrer und die Straße mit Hilfe von Kameras. Zhang et al. [14] stellen hierzu eine Applikation mit der Verbindung eines Farb- und Tiefenbildes vor. Mit Hilfe einer Microsoft Kinect werden sowohl die Kopfpose, als auch die Augenstatus bestimmt. Um das System robuster zu gestalten, wird neben dem Farbbild, auch das Tiefenbild berechnet. Mit der „CarSafe App“ entwickelten You et al. [9] ein visuelles System zur Überwachung des Fahrers und der Straße. Hierfür genügt ein aktuelles Smartphone. Die App deckt hierbei neben der Müdigkeitserkennung auch andere Gefahrensituationen (beispielsweise zu dichtes Auffahren) ab. Es wird eine Analyse des Fahrers (Kopfpose und Augenstatus), sowie seiner Fahrweise kombiniert und entsprechend gewarnt. Bergasa et al. [15] extrahierten aus dem Bild einer Infrarot Kamera mehrere Features, wie beispielsweise den prozentualen Anteil von geschlossenen Augen (Percent eye closure, PERCLOS). Mit dieser Technik erreichten sie bei der Erkennung von Übermüdung eine nahezu hundertprozentige Erfolgsrate. Kamerabasierte Systeme sind angenehm für den Fahrer, da er keine weitere Hardware (Sensoren) installieren muss. Jedoch ist eine Kamera optischen Grenzen unterworfen. Dies erschwert den Einsatz bei Nacht oder schlechtem Wetter. Für eine Müdigkeitserkennung mit Smartphone, aber ohne Kameraeinsatz, könnte beispielsweise die App „V-Sense“ [8] genutzt werden, da sie lediglich eingebaute Sensoren nutzt.

Bundele und Banerjee [16] zeigten, dass Müdigkeit über die elektrodermale Hautreaktion und Pulsoxymetrie erkannt werden kann. Diese wird auch als galvanische Hautreaktion (GSR) bezeichnet und misst die Hautleitfähigkeit, welche wiederum mit der Schweißproduktion zusammenhängt. Bei der Pulsoxymetrie kann, durch ein optisches Verfahren, die Sauerstoffsättigung des Blutes gemessen werden. In diesem Fall bedeutet eine geringere Sättigung ein erhöhtes Müdigkeitsgefühl. Diese Werte werden durch Körpersensoren ermittelt und werden von einem Multi Layer Perceptron (MLP) klassifiziert. Interessant ist zudem der Einsatz von sogenannten Smart-Clothes (E-textiles), welche die Sensoren in der Kleidung eingearbeitet haben und einen non-invasiven Ansatz darstellen.

Park et al. [17] beschränken sich in ihrer Arbeit auf die Analyse der Pulswelle durch Photoplethysmography (PPG), mit einem eingebauten Sensor am Lenkrad. Dies stellt schon ein größeren Eingriff in die Umgebung des Fahrzeugs dar, als es beim vorherigen Verfahren, der Fall war. Die Daten der PPG werden mit einer Support Vector Maschine (SVM) eingeordnet. Es zeigte sich, dass die Ausschlagshöhe des Pules ein gutes Mittel für die Erkennung von Müdigkeit darstellt. Um die Ergebnisse zu verbessern, wurde die vorgestellte Software mit einem zuvor entwickelten, CV-basierten System zur Müdgikeitserkennung gekoppelt.

Zhang et al. [18] befassten sich ebenfalls mit der Überwachung von Herzfunktionen, jedoch mit Hilfe eines EKGs. Sie fanden heraus, dass die Wavelet Packet Energie in bestimmten Frequenzbereichen auf eine Veränderung des QRS-Komplexes hinweist. Dieser kann wiederum als Indiz für einfallende Müdigkeit genutzt werden. Durch diesen Schritt erhöht sich die Geschwindigkeit und die Genauigkeit der Erkennung. In Verbindung mit der Wavelet Entropie kann eine trainierte SVM nahezu 100% erreichen. Rogado et al. [19] nutzen ebenfalls Daten aus dem EKG, um daraus die Herzfrequenzvariabilität (HFR) zu berechnen und drohendes Einschlafen zu erkennen. Ähnliches Verhalten untersuchten Vicente et al. [20].

Khushaba et al. [21] versuchten das EKG Signal mit dem EEG Signal zu koppeln, um die Qualität der Erkennung zu verbessern. Sie verglichen die Erfolgsrate von EKG und EEG, sowie EOG und EEG. Sie nutzen hierfür einen „fuzzy wavelet“ basierten Algorithmus und zeigten, dass das EEG, sowie die Kombination des EEG mit EKG oder EOG gute Ergebnisse liefert. Johnson et. al [22] führten Versuche mit einem EEG und EOG durch. Mit der Linearen Diskriminanzanalyse (LDA) zur Klassifizierung fanden sie heraus, dass das EEG alleine ausreicht und das EOG nicht benötigt wird. Hierfür wurden zwei parallele Studien durchgeführt. Wilson et al. [23] nutzten einen ähnlichen Ansatz, betrachteten jedoch von vornherein nur das EGG und versuchten die Klassifizierung mit einem Künstlichen Neuronalen Netzwerk (KNN) durchzuführen. Sie hielten das KNN durchaus für geeignet, konnten aber kein brauchbares Ergebnis erzielen. Zu einem vergleichbaren Ergebnis kamen Kahlifa et al. [24]. Anders Subasi und Abdulhamit [25], sie konnten mit einer diskreten Wavelet-Transformation und einem KNN ein gutes Resultat erzielen. Sie konnten die Zustände Aufmerksam, Schläfrig (drowsy) und Schlafend mit jeweils über 90 prozentiger Erfolgsrate erkennen. Vuckovic et al. [26] nutzen ebenfalls ein KNN und machten hierzu verschiedene Versuche zum besten Algorithmus zur Erstellung des Netzes. Der Learning Vector Quantization lieferte bessere Ergebnisse, als der Widrow-Hoff Algorithmus und die Levenberg–Marquardt Regel. Im Vergleich zur Klassifizierung von EEG Experten, wurde mit ihrem System eine über 90% Übereinstimmung erreicht. Murthy und Khan [27] verbanden EEG Signale mit einer CV-Technik zur Augenliderkennung und nutzten hierfür ebenfalls ein KNN. Mit Hilfe dieser multimodalen Daten, konnten ebenfalls drei Status erkannt werden: Wach, Schläfrig und Schlafend. Huang et al. [28] nutzten ein Hidden Markov Model (HMM) für die Klassifizierung. Auch dieser Machine Learning Algorithmus ist nach eigener Aussage geeignet, um aus einem EEG-Signal, passende Schlüsse für die Müdigkeitserkennung zu ziehen. Lin et al. [29] nutzten die Unabhängige Komponenten Analyse (UKA) und Lineare Regression (LR) und konnten zeigen, dass hiermit eine Erkennungsrate von bis zu 88% möglich ist.

Die Vielzahl an unterschiedlichen Arbeiten zum Thema Müdigkeitserkennung zeigt die Bandbreite der möglichen Umsetzungen und Kombinationen. Im kommenden Kapitel werden diese Analysiert und Verglichen.

4 Vergleich / Analyse

Nach der Vorstellung der Arbeiten zu Müdigkeitserkennungssystemen, werden diese nun bewertet und verglichen. Die Forschungsergebnisse werden auf Genauigkeit und Störanfälligkeit, sowie Komfort und Portabilität geprüft.
Die beiden ersten Kriterien liegen für ein sicherheitsrelevantes System auf der Hand. Es muss fehlerfrei und robust arbeiten. Fehlender Komfort hingegen ist kein Muss, führt jedoch zu einer geringeren Nutzer-Akzeptanz und verfälscht unter Umständen das Ergebnis, da der Fahrer vom Messsystem abgelenkt werden kann. Um die Anwendung dann in möglichst vielen Szenarien und Umgebungen einzusetzen, ist es wichtig, eine Lösung zu finden, die ohne größeren Aufwand portiert werden kann.

Systeme die das Fahrverhalten analysieren und daraus Rückschlüsse auf den Wachheitsgrad des Fahrers ziehen, sind in der Praxis weit verbreitet. Beispielsweise nutzen Systeme von Daimler [3] oder Volkswagen [10] dieses Verfahren. Die Sensoren (Lenkbewegung, Spurhaltesensoren, Geschwindigkeit usw.) sind entweder fest im Fahrzeug verbaut oder werden per Computer errechnet. Fest verbaute Sensoren müssen auf das jeweilige Fahrzeug abgestimmt werden und sind nicht portierbar. Eingebaute Systeme lassen sich also nicht in einem Fahrzeug oder Simulator nutzen. Für den Fahrer sind diese Systeme jedoch sehr angenehm in der Nutzung, da er vom System nur dann etwas mitbekommt, wenn er vor Müdigkeit gewarnt wird. Da diese Systeme von vielen Automobilherstellern verwendet werden, lässt sich annehmen, dass sie in der Praxis zuverlässig funktionieren. Hierzu konnten jedoch keine Studien gefunden werden, da die Automobilhersteller ihre Forschungsergebnisse nicht veröffentlichen. Dennoch finden sich im Internet mehrere Forenbeiträge, in denen die Nutzer fragen, wie die Pausenempfehlung funktioniert bzw. warum es zu Fehlalarmen kommt[1]. Dies kann darauf hindeuten, dass die Erkennung nicht für alle Bedingungen (Kurven, Unebenheiten auf der Straße) einwandfrei arbeitet.

Kamerabasierte Systeme erreichen ebenfalls sehr gute Ergebnisse (~100% [15]). Für den Fahrer sind diese Systeme komfortabel, da er, wie auch schon bei der Fahrverhaltensanalyse, keinen direkten Kontakt zu den Sensoren hat. Die Kamera (beispielsweise im Smartphone) muss jedoch so ausgerichtet werden, dass gute Bilder aufgenommen werden können. Auch wenn dies der Fall ist, sind kamerabasierte Systeme aber leicht durch äußere Einflüsse beeinflussbar. Die Errechnung der Kopfpose oder Blinzelrate sind dann nicht mehr möglich und es kann zu Fehlalarmen oder schlimmer, zu übersehener Müdigkeit kommen. Schlechte Lichtverhältnisse, wie tief stehende Sonne oder Dunkelheit verändern das Bild einer normalen Kamera und somit auch unter Umständen auch das Ergebnis der Erkennung. Auch wenn Systeme mit Infrarotkameras diese Probleme vermindern, kann es weiterhin sein, das beispielsweise Augen von (Sonnen-)Brillenträgern nicht richtig erkannt werden können. Die Kopfpose und die Augen können bei You et al. [9] beispielsweise nur erkannt werden, wenn der Fahrer in Fahrtrichtung schaut. Anwendungen mit kleinen oder integrierten Kamerasystemen (beispielsweise im Smartphone) lassen sich sehr gut in verschiedenen Umgebungen nutzen. Je aufwändiger die Kameratechnik ist, desto schwieriger lässt sich das gesamte System auf- und abbauen.

Die meisten Körpersensoren liefern sehr gute Erkennungsraten (~100% [18], ∅ 92,25%), es werden EEG, EKG, EOG und weitere Sensoren für die Datensammlung genutzt. Die Daten sind sofort verfügbar und müssen nicht erst als Folgen von Müdigkeit beobachtet werden. Sie lassen sich nicht bewusst verfälschen und sind in den meisten Fällen eindeutig. In den betrachteten Arbeiten wurden meist sehr wenige und ähnliche Probanden getestet (~30) und es bleibt die Frage, ob die Ergebnisse repräsentativ sind. Körpersensoren sind wegen ihrer invasiven Eingenschaften eher unkomfortabel und daher weniger für den Serienbetrieb geeignet. Um dieses Problem zu lösen existieren bereits verschiedene Ansätze. So stellte beispielsweise der französische Automobilzulieferer Faurecia seinen intelligenten Autositz „Active-Wellness“ vor [2]. Der Sitz ist mit passiven Sensoren ausgestattet und misst ständig den Herzrhythmus, die At-

[1] `http://www.motor-talk.de/forum/muedigkeitserkennung-wie-funktioniert-sie-t4515648.html`
Stand 28.10.2015

[2] `http://www.faurecia.de/node/1780`

mung und weitere biometrische Daten, um bei einfallender Müdigkeit gegenzusteuern. Auch Pulsmesser am Handgelenk werden immer besser, sind dabei nicht größer als eine Uhr und bieten ein vertretbaren Tragekomfort. Ein weiterer Vorteil der Lösung ist die universelle Einsetzbarkeit. Diese Systeme sind nicht auf den Fahrzeugbereich beschränkt. Sie können auch in anderen Gefahrenbereichen, beispielsweise in einem Atomkraftwerk, eingesetzt werden. Aufgrund der hohen Genauigkeit von EEG, EKG etc. sind diese Lösungen zur Verbesserung oder Validierung anderer Systeme in der Testphase nützlich. Bei der Datenaufbereitung ist zu beachten, dass Störungen und Rauschen entfernt werden. Für die Klassifizierung wurde in vielen Fällen eine Wavelet Transformation [18][21][25] und ein KNN [23] - [27] verwendet. Gelingt es zudem den invasiven Charakter gering zu halten oder gar ganz zu vermeiden, sind Systeme mit Körpersensoren auch für den Produktiveinsatz geeignet. Es bleibt zu zeigen, dass Genauigkeit und Tragekomfort im richtigen Verhältnis stehen. Hat sich ein System in der Simulationsumgebung bewährt, sollte das System in einem realen Fahrzeug getestet werden können.

Wie gesehen, haben die verschiedenen Ansätze zur Müdigkeitserkennung Stärken und Schwächen. Die Analyse ist in Tabelle 1 noch einmal dargestellt. Das Netzdiagramm in Abbildung 3 zeigt die Tendenzen der einzelnen Ansätze im Durchschnitt.
Systeme mit Körpersensoren zeigten die besten Ergebnisse im Vergleich zu kamera- oder verhaltensbasierten Lösungen. Jedoch ist allen Systemen, unter idealen Bedingungen, eine hohe Genauigkeit gemein. Die Unabhängigkeit von äußeren Einflüssen (Licht- und Straßenverhältnisse) macht Körpersensoren zu einer zuverlässigen Lösung. Wobei kamerabasierte Ansätze am anfälligsten für äußere Störfaktoren sind. Systeme für die Analyse des Fahrverhaltens sind für den Fahrer am angenehmsten, da diese meist im Fahrzeug verbaut sind und keine zusätzliche

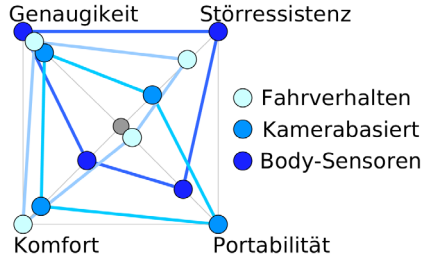

Abbildung 3: Netzdiagramm: Die drei betrachteten Ansätze bezogen auf Genauigkeit, Störanfälligkeit, Komfort und Portabilität. Die Abbildung zeigt eine Zusammenfassung der analysierten Arbeiten.

Hardware notwendig ist. Bei kamerabasierten Systemen kann ein falscher Aufbau zu Problemen führen. Körpersensoren sind aufgrund der invasiven Eigenschaften und zusätzlicher Hardware (Laptop, Smartphone) am unkomfortabelsten. Fest verbaute Sensoren, wie sie für die Analyse des Fahrverhaltens genutzt werden, sind nicht ausbaubar und nicht portabel. Systeme mit Kameras und Körpersensoren setzen sich meisten aus Sensoren und Computer zusammen, diese sind beide portierbar, wobei Körpersensoren meist einen komplizierteren Aufbau haben.
Systeme mit Körpersensoren liefern in zwei von vier Kategorien beste Ergebnisse und werden daher im folgenden Kapitel für die Entwicklung eines System zur Müdigkeitserkennung weiterverfolgt. Eine Komfortsteigerung und die Erhöhung der Portabilität werden Ziele des Konzepts sein.

5 Portables System zur Müdigkeitserkennung mit Körpersensoren

Wie im vorherigen Absatz gesehen, existieren sehr viele verschiedene Lösungen zur Müdigkeitserkennung in Fahrzeugen. Aufgrund der beschriebene Vorteile von Körpersensoren, soll das zu entwickelnde System mit eben diesen arbeiten. Das

Tabelle 1: Vergleichstabelle

Ref.	Sensor	Klassifikator	Merkmal	Genauigkeit (Testpersonen)
		Fahrverhaltensanalyse		
	Keine Daten vorhanden			
		Kamerabasiert		
[14]	Kinect	SVM	Kopfpose, Augenstatus	93% (30)
[9]	Smartphone Kamera	OpenCV Funktionen	Kopfpose, Augenstatus	68% / 88%
[15]	Infrarot Kamera	-	PERCLOS	~100%
		Körpersensorbasiert		
[16]	GSR	KNN	Sauerstoffsättigung	93,17%
[17]	PPG	SVM	Puls-Ausschlagshöhe	80% (5)
[18]	EKG	SVM	QRS-Komplex	~100%
[19]	EKG	-	HFR	-
[20]	EKG	LDA	HFR	93%
[21]	EKG, EEG, EOG	LDA	HFR	97% (31)
[22]	EEG, EOG	LDA	-	- (160)
[23]	EEG	KNN	-	-
[24]	EEG	KNN	-	-
[25]	EEG	KNN	-	93% (30)
[26]	EEG	KNN	-	94% (17)
[28]	EEG	HMM	-	-
[29]	EEG	UKA, LR	-	88,2% (16)
		Mischformen		
[27]	EEG, Kamera	KNN	-	-

vorgeschlagene System soll Genauigkeit, Tragekomfort und Portabilität maximieren (Siehe Abb. 4). Hierfür muss zum einen das Problem des invasiven Körpersensoren gelöst werden. Zum anderen sollte das ganze System leicht aus- und eingebaut werden können und ohne größeren Konfigurationsaufwand funktionieren. Das vorgestellte Konzept soll keine komplette Neuentwicklung sein, sondern erweitert die analysierten Forschungen. Sinnvolle Ansätze werden aufgegriffen und evaluiert. Die Schwerpunkte Komfort und Portabilität wurde bisher nicht ausreichend berücksichtigt, dies soll im Rahmen einer späteren Umsetzung getan werden.

EEG und EKG stehen in der Simulationsumgebung der Reutlingen University zur Verfügung. Bei den eingesetzten Sensoren, soll insbesondere auf den Tragekomfort geachtet werden. Im besten Fall nimmt der Fahrer die Sensoren nicht wahr und wird durch diese nicht abgelenkt. Das EEG scheidet bei dieser Betrachtung von vorne herein aus und soll während der Entwicklung zur Verfeinerung bzw. Validierung genutzt werden. Im Bereich des EKGs existieren Lösungen mit höheren Tragekomfort. Diese liefern jedoch unter Umständen weniger Daten oder solche mit schlechterer Qualität (beispielsweise erhöhtes Rauschen). Daher muss gezeigt werden, dass die Genauigkeit hoch und die Fehlerrate möglichst niedrig

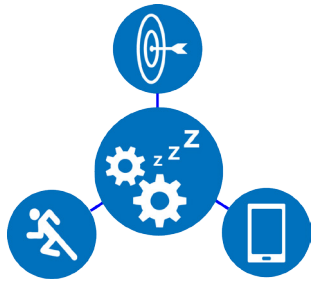

Abbildung 4: Schwerpunkte des Systems zur Müdigkeitserkennung (Im Uhrzeigersinn): Genauigkeit, Portabilität und Tragekomfort

bleibt. Darum wird eine Umsetzung mit einem EKG Brustband (Siehe Kapitel 6.2) vorgeschlagen, welches bei der Entwicklung von einem EEG validiert wird. Ein Austausch des Brustbandes durch einen Pulsmesser oder eine Smartwatch ist vorstellbar.

Um im Fahrzeug verbaute Körpersensoren zu simulieren, werden die Sensordaten in der Simulationsumgebung von einem virtuellen Steuergerät via Controller Area Network-Bus (CAN-Bus) über ein Interface an die Anwendung übertragen. Die Daten können aber auch direkt, ohne Simulator, übertragen werden. Dennoch sind Feldversuche für abschließende Tests unumgänglich und sei es nur, um zu Prüfen, ob es in einer realen Testfahrt zu Fehlalarmen kommt (Stress, Beschleunigung oder Ähnliches). Darum muss die Anwendung sowohl im Simulator, als auch in einem realen Fahrzeug funktionieren und ohne großen Aufwand portiert werden können. Wenn dies gelingt, kann die Anwendung sowohl im Simulator der Reutlingen University, einem realen Fahrzeug oder einem anderen Simulator genutzt werden, um dort etablierte Systeme zur Müdigkeitserkennung zu ergänzen oder zu validieren. Das vorgestellte System muss demnach mit möglichst wenig Hardware auskommen und die Software leicht auf andere Geräte portieren lassen. Im einfachsten Fall genügt ein Laptop oder Smartphone. Auch hier wäre eine Erweiterung mit einer Smartwatch denkbar.

Um das System auch zur Verbesserung oder Validierung anderer Systeme zur Müdigkeitserkennung zu nutzen, werden öffentliche Schnittstellen definiert und ein sauberes Logging der Daten implementiert. Die Daten der Überwachung sollen mit eindeutigem Zeitstempel versehen werden. So können sie mit anderen Daten, wie Videoaufzeichnung oder Ergebnissen anderer Systeme, zu einem späteren Zeitpunkt verglichen werden.

Erkennt das System eine drohende Müdigkeit, soll es den Fahrer über ein optisches oder akustisches Signal warnen. Das System soll mehrere Warnstufen kennen und je nach Müdigkeitsgrad reagieren.

Die Anforderungen eines portablen Systems zur Müdigkeitserkennung mit Körpersensoren waren Thema des vergangenen Kapitels. Im folgenden wird das weitere Vorgehen, sowie Voraussetzungen zur Umsetzung beschrieben.

6 Evaluationsplan

Um das System zu implementieren, müssen die notwendigen Schritte ausgearbeitet werden (Siehe Abb. 5). Bausteine sind zu Beginn die Datenaggregation der EKG- und EEG-Sensoren. Diese Daten müssen in der Simulationsumgebung via CAN-Bus bzw. später im realen Fahrzeug direkt Übertragen werden. Die Datenaufbereitung und Klassifizierung sind Aufgaben der Anwendung. Die Rückmeldung erkannter Müdigkeit wird ebenfalls von der Anwendung angestoßen. Bei der Evaluation müssen die Anforderungen Genauigkeit, Tragekomfort und Portierbarkeit beachtet werden. In den folgenden Absätzen werden die einzelnen Schritte näher beschrieben, sowie ein Szenario zur Erhebung von Testdaten vorgestellt.

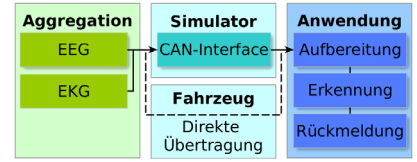

Abbildung 5: Der Ablauf der Datenströme und Aufgaben des Systems. Aggregation der Sensordaten, Datenübertragung im Simulator oder Fahrzeug, Verarbeitung der Daten in der Anwendung

6.1 Datenaggregation

Signale aus EEG und EKG sind vielschichtig und müssen auf ihre Tauglichkeit geprüft werden. Dies kann im ersten Schritt unabhängig von Simulationsumfeld geschehen. Latenz oder Störsignale müssen beachtet und möglichst eliminiert werden. Die Aufbereitung der Signale erfolgt zu einem späteren Zeitpunkt in der Anwendung.

Für das System wird ein EEG Epoc Emotiv [3] und ein Brustband EKG Zephyr Bioharnes [4] eingesetzt (Siehe Abb. 6). Das EEG ist in der Simulationumgebung der Reutlingen University bereits einsatzbereit. Die Anbindung des EKG Brustbands muss noch implementiert werden.

Abbildung 6: Links: EKG Brustband Zephyr Bioharnes; Rechts: EEG Epoc Emotiv

[3] https://emotiv.com/epoc.php
[4] http://www.zephyranywhere.com/products/bioharness-3

Sind die ersten Versuche mit den Sensoren erfolgreich, werden sie in die Simulationsumgebung integriert.

6.2 Übertragung im Simulator

Die erfassten Sensor-Daten müssen an den Simulator angeschlossen werden, um sie dann an die Anwendung übertragen zu können.
Die Simulationsumgebung der Reutlingen University ermöglicht es, Testfahrten unter möglichst realistischen Bedingungen durchzuführen (Siehe Abb. 6.2). Sie ist ausgestattet mit einem Autositz, einem Lenkrad und Pedalen, sowie einer Gangschaltung. Weiterhin bringen drei 48″ Monitore und ein Dolby-Surround-System einen audiovisuellen Eindruck einer Fahrsituation.

Auf technischer Ebene wird das System von drei Computern betrieben. Auf dem Simulations-Computer läuft die vom DFKI Saarbrücken entwickelte Software OpenDS [5]. Hier können mehrere Karten und Konfigurationen eingestellt und getestet werden. Alle Simulationsdaten, werden via TCP/IP an den Daten-Computer gesendet. Dieser stellt die Daten über ein Interface zur Verfügung und loggt diese zusätzlich. Der eingesetzte CAN-Bus simuliert die Kommunikation mit dem Steuergerät eines realen Fahrzeugs. Der Anwendungs-Computer kann die CAN Daten über eine Schnittstelle empfangen und seine eigentliche Arbeit verrichten. Für Ein- und Ausgaben der Anwendung, befindet sich ein Touchscreen neben dem Lenkrad. Das System zur Müdigkeitserkennung wird auf dem Anwendungs-Computer ausgeführt.

6.3 Ablauf der Anwendung

Die übertragenen Daten müssen aufbereitet werden, sodass der Klassifizierer trainiert und getestet werden kann. Das kontinuierliche Signal muss im ersten Schritt in ein diskretes Signal umgewandelt werden. Dann können die Werte weiterverarbeitet werden, um aussagekräftige Merkmale für die Klas-

[5] http://www.dfki.de/web/aktuelles/aktuelles/cebit2013/opends

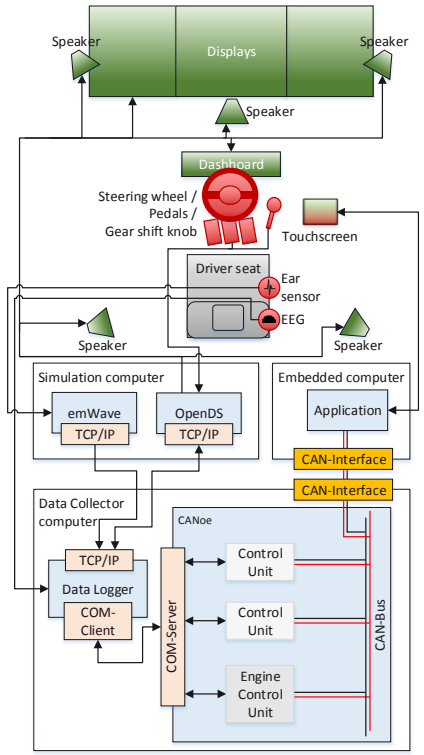

Abbildung 7: Der Aufbau des Simulators der Reutlingen University

sifizierung zu erhalten. In vielen Ansätzen wird die Wavelet-Transformation [30] erfolgreich eingesetzt. Alternativen wie die Fourier- [31] oder Laplace-Transformation [32] sind ebenfalls zu prüfen.
Im nächsten Schritt benötigt man für die Erkennung von Müdigkeit einen Klassifikator bzw. Machine-Learning-Algorithmus. In den betrachteten wissenschaftlichen Arbeiten wurde häufig auf KNNs [33] oder SVMs [34] vertraut. Weitere Machine-Learning-Ansätze, wie das Hidden-Markov-Modell oder Lineare Regression wären mögliche Alternativen.

Wird eine Übermüdung erkannt, wird der Fahrer durch ein optisches oder akustisches Signal aufgefordert eine Pause einzulegen.

6.4 *Testszenario*

Das portable System zu Müdigkeitserkennung wird am Fahrsimulator der Reutlingen University entwickelt und getestet. Die Simulation kann jedoch nur ein Modell der Wirklichkeit sein und liefert nur eingeschränkte Ergebnisse. So konnten Blana et al. [35] zeigen, dass sich das Fahrverhalten bei höheren Geschwindigkeiten im echten Straßenverkehr und im Simulationsumfeld unterscheidet. Engstrom et al. [36] konnten ebenfalls Unterschiede feststellen, zeigten jedoch auch, das Tests im Simulator dennoch valide Ergebnisse liefern können.

Um Daten mit einfallender Müdigkeit zu erhalten, muss ein passendes Szenario im Simulator erstellt werden. Es gilt beim Versuchsaufbau eine möglichst große Chance auf Sekundenschlaf bei den Probanden zu provozieren. Horne und Reyner [37] legten nahe, dass die meisten Unfälle im Zusammenhang mit Schlaf in Großbritannien zwischen 02:00 - 06:00 und 14:00 - 16:00 passierten. Weiterhin lässt sich beobachten, dass Personen die 24 Stunden gar nicht oder nicht ausreichend (< 6h) geschlafen haben, deutlich anfälliger für Sekundenschlaf sind [38]. Ein weiterer Faktor ist das Testszenario selbst. Es sollte möglichst gut erkennbar machen, ob der Proband gerade Fahrfehler macht und diese mit seinem Wachheitsgrad zusammenhängen. Weiterhin kann auch eine „langweilige Teststrecke“ eine schnellere Ermüdung begünstigen. Langes geradeaus fahren mit wenig Abwechslung sind langweilig und könnten mit einer Spurhalte-Aufgabe gekoppelt werden, sodass der Probanden über die ganze Zeit gefordert ist. Auch die Länge der Fahrt spielt eine Rolle, je länger die Fahraufgabe dauert, desto größer die Chance auf das Eintreten einer einfallenden Müdigkeit.

Vorstellbar ist somit folgender Ablauf: 1) 10min Einführung und Fahrsimulator ausprobieren. 2) 10min Straßenverkehr mit anderen Verkehrsteilnehmern unter Beachtung der Straßenverkehrsregeln. 3) 20min Gerade aus fahren und Spur halten 4) 10min Wiederholung von Schritt 2. 5) 10min Befragung und Selbsteinschätzung. Im besten Fall lassen sich bei der Spurhalte-Aufgabe gegen Ende mehr Fahrfehler beobachten. Weiterhin sollte der Fahrer in Schritt 4) ein Unterschied zu Schritt 2) feststellen lassen.

In den vergangenen Abschnitten wurde die Vorgehensweise und weitere Schritte näher erläutert. Im kommenden Kapitel werden die Ergebnisse zusammengefasst und weitere Schritte beschrieben.

7 Fazit und Ausblick

Die Erkennung von Müdigkeit stellt einen wichtigen Beitrag zu Sicherheit im Straßenverkehr dar. Schwere Unfälle können mit einem solchen System verhindert werden. Dazu ist es notwendig, möglichst früh und präzise eine drohende Müdigkeit zu erkennen. Für die Erkennung existieren verschiedene Ansätze, die sich in kamerabasierte, Fahrverhaltenanalysierende und Körpersensorbasierte Systeme einteilen lassen.

Im evaluierten Stand der Technik lieferten Systeme mit Körpersensoren sehr gute Ergebnisse, zeigen aber im Bereich Komfort und Portabilität schwächen. Sie sind dennoch robust und den beiden anderen Ansätzen mit Kamera und Fahranalyse überlegen. Körpersensoren liegen jedoch meist direkt am Körper an und können den Fahrer während der Fahrt stören. Das EKG-Brustband stellt vermutlich einen guten Kompromiss von Tragekomfort und Genauigkeit dar. Ob die Genauigkeit des Sensors für eine frühzeitige Erkennung ausreicht, wird mit Hilfe eines EEGs belegt werden müssen. Die Entwicklung des Systems wird im Simulatorumfeld der Reutlingen University durchgeführt. Dennoch muss das System in einem echten Fahrzeug einsetzbar sein und somit ohne größeren Aufwand portierbar sein. Zur Entwicklung des Systems werden zudem geeignete Testdaten benötigt. Das aufgezeigte Konzept ist durch ähnliche Arbeiten aus diesem Bereich fundiert und senkt damit das Risiko einer Sackgasse. Viele iterative Ausbaustufen ermöglichen es, frühzeitig Hürden zu erkennen und Probleme zu lösen. Die Umsetzung des Konzepts erfolgt im Rahmen des IoT-Labs [6] der Reutlingen University.

Das zu entwickelnde System soll zur Verbesserung oder Validierung anderer Systemen zur Müdigkeitserkennung eingesetzt werden. Langfristig ist eine Version für den Produktivbetrieb denkbar, aber nur wenn auf non-invasive Körpersensoren zurückgegriffen werden kann.
Für die Umsetzung sind folgende Schritte vorgesehen. Einbau der Sensoren in die Simulationsumgebung der Reutlingen University. Die Daten des EKG und EEG müssen aufbereitet werden, um sie zur Klassifizierung, in einen passenden Machine-Learning Algorithmus, geben zu können. Um Testdaten zu erhalten, muss das vorgestellt Szenario durchgeführt werden. Funktioniert die Erkennung im Simulationsumfeld, muss das System Schrittweise verkleinert (Hardware- und Softwareseitig) und vom Simulator entkoppelt werden. Dann können erste Tests in einem echten Fahrzeug stattfinden. Das System soll in der weiteren Entwicklung mit anderen Systemen zu Müdigkeitserkennung verglichen werden und ggf. an einer Kopplung der Anwendungen gearbeitet werden. In der letzten Ausbaustufe, ist vorstellbar, dass das System komplett in einer Smartwatch mit Pulsmesser läuft. Damit wäre die Anwendung mit keinerlei Beeinträchtigung des Fahrers verbunden. Ob die Sensorgenauigkeit und die Rechenleistung der Smartwatch ausreicht, bleibt zu zeigen.

[6] `http://iotlab.reutlingen-university.de`

Literatur

[1] Strategy Analytics. Advanced driver assistance systems forecast - aug 2015. https://www.strategyanalytics.com/access-services/automotive/powertrain-body-chassis-and-safety/market-data/report-detail/advanced-driver-assistance-systems-forecast---aug-2015, 2015. Zugriff: 2015-10-28.

[2] Xavier Mosquet, Michelle Andersen, and Aakash Arora. A roadmap to safer driving through advanced driver assistance systems. https://www.bcgperspectives.com/Images/MEMA-BCG-A-Roadmap-to-Safer-Driving-Sep-2015.pdf, 2015. Zugriff: 2015-10-28.

[3] Daimler AG. Attention assist, 2008. Available at http://media.daimler.com/dcmedia/0-921-658892-49-1147698-1-0-0-0-0-1-11702-854934-0-1-0-0-0-0-0.html, Zugriff: 2015-08-13.

[4] Claudia Evers. Unterschätzte Risikofaktoren Übermüdung und ablenkung als ursachen für schwere lkw-unfälle. http://www.dvr.de/presse/seminare/904_20.htm, 2008. Zugriff: 2015-10-20.

[5] National Sleep Foundation. Drivers Beware: getting enough sleep can save your life this memorial day. http://us1.campaign-archive.com/?u=72c7dac36ef8bcb0852893d7c&id=56d88442b5, 2010. Zugriff: 2015-10-20.

[6] Klaus Kompaß. Fahrerassistenzsysteme der zukunft – auf dem weg zum autonomen pkw? In Volker Schindler, editor, *Forschung für das Auto von Morgen*, pages 261–285. Springer Berlin Heidelberg, 2008.

[7] Eduardo Bertoldi and Lucia Filgueiras. Multimodal advanced driver assistance systems: An overview. In *Proceedings of the 2Nd International Workshop on Multimodal Interfaces for Automotive Applications*, MIAA '10, pages 2–5, New York, NY, USA, 2010. ACM.

[8] Dongyao Chen, Kyong-Tak Cho, Sihui Han, Zhizhuo Jin, and Kang G. Shin. Invisible sensing of vehicle steering with smartphones. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '15, pages 1–13, New York, NY, USA, 2015. ACM.

[9] Chuang-Wen You, Nicholas D. Lane, Fanglin Chen, Rui Wang, Zhenyu Chen, Thomas J. Bao, Martha Montes-de Oca, Yuting Cheng, Mu Lin, Lorenzo Torresani, and Andrew T. Campbell. Carsafe app: Alerting drowsy and distracted drivers using dual cameras on smartphones. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '13, pages 13–26, New York, NY, USA, 2013. ACM.

[10] Robert Bosch GmbH. Bosch driver drowsiness detection, 2012. Available at http://www.bosch-presse.de/presseforum/details.htm?txtID=5037, Zugriff: 2015-08-13.

[11] Joan Santamaria and Keith H. Chiappa. The eeg of drowsiness in normal adults. *Journal of Clinical Neurophysiology, Volume 4*, pages 327–382, 1987.

[12] Deutscher Verkehrssicherheitsrat e.V. Müdigkeit im straßenverkehr, 2009. Available at http://www.dvr.de/dvr/vorstandsbeschluesse/vm-ft_muedigkeit.htm, Zugriff: 2015-08-13.

[13] Arun Sahayadhas, Kenneth Sundaraj, and Murugappan Murugappan. Detecting driver drowsiness based on sensors: A review. *Sensors*, 12(12):16937, 2012.

[14] Liyan Zhang, Fan Liu, and Jinhui Tang. Real-time system for driver fatigue detection by rgb-d camera. *ACM Trans. Intell. Syst. Technol.*, 6(2):22:1–22:17, March 2015.

[15] Luis M. Bergasa, Jesus Nuevo, Miguel A. Sotelo, Rafael Barea, and Maria E. Lopez. Real-time system for monitoring driver vigilance. *Intelligent Transportation Systems, IEEE Transactions on*, 7(1):63–77, March 2006.

[16] Mahesh M. Bundele and Rahul Banerjee. Detection of fatigue of vehicular driver using skin conductance and oximetry pulse: A neural network approach. In *Proceedings of the 11th International Conference on Information Integration and Web-based Applications &Amp; Services*, iiWAS '09, pages 739–744, New York, NY, USA, 2009. ACM.

[17] Hanbit Park, Seungwon Oh, and Minsoo Hahn. Drowsy driving detection based on human pulse wave by photoplethysmography signal processing. In *Proceedings of the 3rd International Universal Communication Symposium*, IUCS '09, pages 89–92, New York, NY, USA, 2009. ACM.

[18] Aihua Zhang and Fenghua Liu. Drowsiness detection based on wavelet analysis of ecg and pulse signals. In *Biomedical Engineering and Informatics (BMEI), 2012 5th International Conference on*, pages 491–495, Oct 2012.

[19] E. Rogado, J.L. Garcia, Rafael Barea, Luis M. Bergasa, and Elena Lopez. Driver fatigue detection system. In *Robotics and Biomimetics, 2008. ROBIO 2008. IEEE International Conference on*, pages 1105–1110, Feb 2009.

[20] Jose Vicente, Pablo Laguna, Ariadna Bartra, and Raquel Bailon. Detection of driver's drowsiness by means of hrv analysis. In *Computing in Cardiology, 2011*, pages 89–92, Sept 2011.

[21] Rami N. Khushaba, Sarath Kodagoda, Sara Lal, and Gamini Dissanayake. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *Biomedical Engineering, IEEE Transactions on*, 58(1):121–131, Jan 2011.

[22] Robin R. Johnson, Djordje P. Popovic, Richard E. Olmstead, Maja Stikic, Daniel J. Levendowski, and Chris Berka. Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biological psychology*, 87(2):241–250, May 2011.

[23] Beth J. Wilson and Thomas D. Bracewell. Alertness monitor using neural networks for eeg analysis. In *Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop*, volume 2, pages 814–820 vol.2, 2000.

[24] Khaled B. Khalifa, Mohamed H. Bedoui, R. Raytchev, and Mohamed Dogui. A portable device for alertness detection. In *Microtechnologies in Medicine and Biology, 1st Annual International, Conference On. 2000*, pages 584–586, 2000.

[25] Abdulhamit Subasi. Automatic recognition of alertness level from eeg by using neural network and wavelet coefficients. *Expert Syst. Appl.*, 28(4):701–711, May 2005.

[26] Aleksandra Vuckovic, Vlada Radivojevic, Andrew C.N. Chen, and Dejan Popovic. Automatic recognition of alertness and drowsiness from {EEG} by an

artificial neural network. *Medical Engineering & Physics*, 24(5):349 – 360, 2002.

[27] Keshava Murthy and Zaved Ahmed Khan. Smart alert system for driver drowsiness using eeg and eyelid movements. In *Middle-East Journal of Scientific Research 14*, pages 610–619. IDOSI Publications, 2013.

[28] Ruey S. Huang, Chung .J. Kuo, Ling-Ling Tsai, and Oscal T.C. Chen. Eeg pattern recognition-arousal states detection and classification. In *Neural Networks, 1996., IEEE International Conference on*, volume 2, pages 641–646 vol.2, Jun 1996.

[29] Chin teng Lin, Ruei cheng Wu, Sheng fu Liang, Wen hung Chao, Yu jie Chen, and Tzyy ping Jung. Eeg-based drowsiness estimation for safety driving using independent component analysis. *IEEE Trans. Circuits Syst. I, Reg. Papers*, pages 2726–2738, 2005.

[30] Charles K. Chui. *An Introduction to Wavelets*. Academic Press Professional, Inc., San Diego, CA, USA, 1992.

[31] Chandrasekharan K. Bochner S. *Fourier Transforms*. Princeton University Press, 1949.

[32] David Vernon Widder. *The Laplace Transform*. Princeton University Press, 1941.

[33] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[34] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

[35] Evi Blana and John Golias. Differences between vehicle lateral displacement on the road and in a fixed-base simulator. *Human Factors*, 44(2):303–313, 2002.

[36] Johan Engstrom, Emma Johansson, and Joakim Ostlund. Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(2):97–120, March 2005.

[37] Jim Horne and Louise Reyner. Vehicle accidents related to sleep: a review. *Occupational and Environmental Medicine*, pages 289–294, May 1999.

[38] Robert Peters, Esther Wagner, Elizabeth Alicandri, Jean Fox, Maria L. Thomas, David R. Thorne, Helen C. Sing, and Sharon M. Balwinski. Effects of partial and total sleep deprivation on driving performance. *Public Roads*, pages 2–6, May 1999.

Simulationsansatz für die Entwicklung von kognitiv technischen Komponenten zur Bewegungswahrnehmung *

David Randler
Reutlingen University
David.Randler@Student.Reutlingen-University.DE

Abstract

Die automatisierte Analyse von menschlichen Bewegungen ermöglicht die Erschließung neuer Anwendungsfelder. Zugleich ist ein erfolgreicher Trend, insbesondere in der Automobilindustrie, die quasi realistische Simulation von großen Sensordatenmengen zur Optimierung von Systemkomponenten. Dieser Simulationsansatz ermöglicht die flexible Analyse von Systembestandteilen und Algorithmen, ohne an reale Begebenheiten gebunden zu sein. Auf Grund dessen untersucht dieser Artikel das Potential der grafischen Simulation menschlicher Bewegungen zur Systementwicklung, wobei insbesondere auf den Nutzen für die Entwicklung komplexer Sehfunktionen der Bildverarbeitung eingegangen wird. Es wird herausgearbeitet, welches Vorgehen nötig ist, um menschliche Bewegungen realistisch zu simulieren und welche Probleme dabei auftreten können. Anhand der Bestimmung von Kriterien werden mögliche Simulationsumgebungen im Kontext der Bewegungssimulationen analysiert. Außerdem wird im Detail erläutert durch welche Anwendungen der Simulationen die Bildverarbeitung profitieren kann.

Schlüsselwörter

Motion Capture, Simulation, Retargeting, Human Character Animation, Domain Adaption, Computer Vision

CR-Kategorien

I.3.7 [**Computer Graphics**]: Three-Dimensional Graphics and Realism — Animation;

1 Einleitung

Menschliche Bewegung spielt eine zentrale Rolle in unserem Handeln. Die Änderung des Gesichtsausdrucks als Reaktion auf eine andere Person oder die Bewegung der Beine zum Erreichen eines Ziels sind letztlich Bewegungsabläufe unseres Körpers. Aus diesem Grund kann durch die gezielte Analyse von menschlichen Bewegungen eine Vielzahl von Wissen generiert werden. Gesten können Aufschluss über soziale Interaktionen geben, Arm- und Beinbewegungen im Straßenverkehr können auf mögliche Intentionen von Fußgängern und Fahrern hindeuten und Bewegungsabläufe können für die Verbesserung von Ergonomieaspekten eines Produktes genutzt werden. Dies sind nur einige der Anwendungsfelder, für deren Forschung die Analyse menschlicher Bewegung erforderlich und sinnvoll ist. Gerade durch die Komplexität und Vielfalt der menschlichen Bewegungen ist dies jedoch

---

*

Betreuer Hochschule: Prof. Dr.-Ing. Cristóbal Curio
Hochschule Reutlingen
cristobal.curio@Reutlingen-University.de
Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

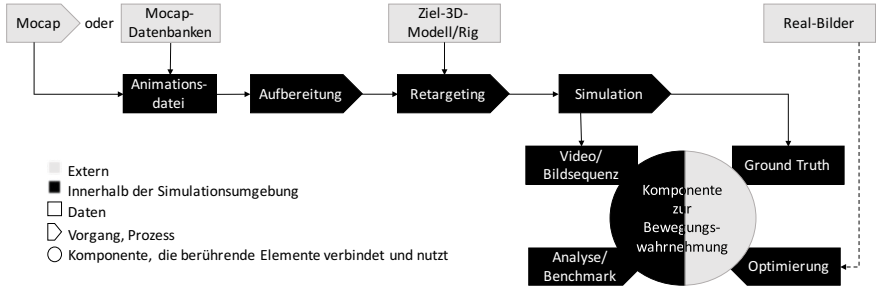

Abbildung 1: Gesamter Ablauf eines Simulationsansatzes für die Entwicklung von kognitiv technischen Komponenten.

sehr schwierig und zeitaufwendig. So müssen für Leistungstests von Algorithmen 3D-Aufnahmen von Körpern erstellt werden, die möglichst variabel, präzise, gut geeignet und repräsentativ sind [3]. Dies durch manuelle Aufnahme und insbesondere Annotation der Referenzdaten (engl. Ground Truth) zu erreichen, ist nur mit erheblichem Aufwand ansatzweise zu schaffen. Ohne diese Daten auszukommen, ist dagegen keine Alternative, da einerseits die Tauglichkeit des Algorithmus nachgewiesen und andererseits möglicherweise eben diese Daten für dessen Training benötigt werden. Insbesondere aktuell vermehrt verwendete Verfahren des maschinellen Lernens wie beispielsweise Regression Forests [5] basieren auf einer sehr großen Anzahl an Trainigsdaten.

Aus dem hohen Aufwand für die manuelle Erstellung von Test- und Trainingsdaten ergibt sich schließlich die Idee, menschliche Bewegungsabläufe zu simulieren. Dies ist ebenfalls schwierig und komplex, erlaubt jedoch eine neue Dimension an Flexibilität und eine automatische Generierung der Referenzdaten. Aus diesem Grund wird im folgenden Abschnitt das Vorgehen für die Simulation von menschlichen Bewegungen herausgearbeitet und es wird detaillierter auf aktuelle und etablierte Ansätze, wichtige Schritte und Probleme eingegangen. Anschließend werden Kriterien für den Aufbau einer Simulationsumgebung aufgestellt und für zwei potentielle Umgebungen evaluiert. Vor einem abschließenden Fazit wird die Anwendung von Simulationen und deren Nutzen in der Bildverarbeitung ausführlich diskutiert.

2 Simulation von menschlichen Bewegungen

Die Simulation hat das Ziel, menschliche Bewegungen auf natürliche Weise in einer virtuellen Umgebung abzubilden. Daraus folgt umgehend, dass Animationen, die lediglich auf Keyframes basieren, nicht geeignet sind. Bei solche Animationen werden Zeitschritt für Zeitschritt die Positionen bestimmter Körper- beziehungsweise Objektelemente positioniert. Dies kann bei Animationsfilmen ausreichen, ist für natürliche menschliche Bewegungsabläufe jedoch kaum anwendbar. Für diese Simulationen wird auf sogenannte Mocaps, das heißt Bewegungsaufnahmen (engl. Motion Capture) zurückgegriffen, die an ein virtuelles Modell gekoppelt werden. Abbildung 1 stellt den gesamten Ablauf des Simulationsansatzes schematisch dar. Die folgenden Abschnitte gehen im Detail auf die Aspekte ein, die bis zur Simulation menschlicher Bewegungen erforderlich sind.

2.1 Mocap

Mocap (Motion Capture) beschreibt nach [20] und [4] das Aufnehmen von Bewegungen eines Individuums und deren Weiterverarbeitung am Computer. Dabei werden die Daten häufig auf einen virtuellen Charakter projiziert. Für die Aufnahme werden verschiedene Systeme verwendet, die sich in ihrer Technologie unterscheiden und entsprechend klassifiziert werden können [20]. Optische Systeme sind sehr verbreitet und nutzen entweder Marker-LEDs, -Reflektoren oder funktionieren ohne jegliche Marker, die am Köper angebracht werden müssen. Ein Beispiel für letzteres ist die *Microsoft Kinect*[1]. Marker-Systeme gelten dagegen als robuster und sind für größere räumliche Gegebenheiten geeignet. Neben optischen Systemen wird versucht durch das Anbringen von Sensoren an Körperteilen deren relative Bewegung zu berechnen. Genutzt werden beispielsweise Goniometer, Magnetsensoren und/oder Trägheitssensoren. Ein aktuelles Beispiel eines solchen Systems stellt der in Abbildung 2 dargestellte *Perception Neuron*[2] dar, welcher Gyroskop, Beschleunigungs- und Magnetsensoren nutzt. Ein Vorteil dieser Systeme ist, dass Verdeckungen im Vergleich zu optischen Systemen keine Rolle spielen.

Die Daten, die durch solche Systeme aufgenommen werden, bestehen für Körperbewegungen meist aus einem Skelett und einer initialen Pose, sowie den Bewegungsdaten für jedes Gelenk [25]. Das Skelett wird durch das Aufnahme-System definiert. Im Allgemeinen wird von einem hierarchischen Modell ausgegangen, bei dem ein Gelenk - oftmals die Hüfte - als Wurzel angesehen wird. Die Definition des Skeletts bestimmt durch die Gelenkpunkte und Knochenlängen die abbildbare Genauigkeit, zum Beispiel ob Fingerbewegungen abgebildet werden können. Trotz der Ähnlichkeit der Daten existieren durch die proprietäre Entwicklung von Mocap-Systemen unterschiedlichste Dateiformate. Durchgesetzt haben sich insbesondere BVH, FBX und teils DAE, ASF/AMC und C3D. Letzteres ist das einzige Format, das statt eines Skeletts lediglich Marker IDs ohne jegliche Zuordnung speichert. Solch ein Format könnte aus diesem Grund auch für Aufnahmen von Gesichtszügen geeignet sein.

[1] http://www.xbox.com/kinect/, 28.09.2015
[2] https://neuronmocap.com/, 28.09.2015

Abbildung 2: Perception Neuron, ein auf Microcontrollern und Sensoren basierendes Mocap-System[2]

2.2 Mocap-Datenbanken

Die inzwischen etablierten Dateiformate erlauben es, neben der eigenen Aufnahme von Bewegungsdaten mit teurer Mocap-Hardware, bestehende Bewegungen einsetzen zu können. So bietet die Carnegie Mellon University eine Datenbank von Bewegungen als ASF/AMC an[3], die jedoch ebenso in BVH konvertiert wurden[4]. Weitere nützliche Quellen sind die ACCAD Open Motions der Ohio State University[5] und die HDM05 Datenbank des Max-

[3] http://mocap.cs.cmu.edu/, 10.09.2015
[4] https://sites.google.com/a/cgspeed.com/cgspeed/motion-capture/cmu-bvh-conversion, 10.09.2015
[5] http://accad.osu.edu/research/mocap/mocap_data.htm, 10.09.2015

Planck-Instituts[6]. Neben diesen Forschungsdatenbanken können Mocap-Daten kommerziell erworben werden. Qualitativ hochwertig sind Daten von den Anbietern Mixamo[7] oder AXYZ Design[8].

2.3 Menschliche Modelle

Abbildung 3: Die Kombination von Mesh und Skelett (grau) führt zu einem Rig.

Abgesehen von den im vorigen Abschnitt beschriebenen Mocap-Daten werden für eine realitätsnahe Simulation von menschlichen Bewegungen virtuelle Menschmodelle benötigt, falls von Körperbewegungen ausgegangen wird. Diese bestehen grundsätzlich aus einem Polygonnetz (Mesh), das die Form des Körpers darstellt [7]. Gewinnen lassen sich solche Strukturen beispielsweise mit einem 3D-Scan. Um Bewegungen abbilden zu können, muss das Mesh an ein Skelett gebunden werden. Unter diesem sogenannten *Rigging*, wird die Zuordnung von Vertices des Meshs zu Knochen des Skeletts verstanden, wobei eine Gewichtung stattfinden kann. So wirkt die Bewegung eines Knochens weniger auf die Änderungen eines Vertex ein als die eines anderen und es wird eine realistischere Deformation erreicht. Oftmals sind Menschmodelle bereits geriggt, beispielsweise bei der Generierung mittels des Tools MakeHuman[9], wie in Abbildung 3 zu sehen, oder bei einem Kauf von Mixamo oder AXYZ Design. Ein manuelles Rigging eines Meshes ist dagegen aufwendig, wenn das Ergebnis eine realistische Verformung des Körpers ohne Artefakte sein soll. Automatisierte Verfahren werden in [21] diskutiert. Im vorliegenden Fall wird von einem bereits geriggten Modell ausgegangen und das Ziel ist, dieses mit den Mocap-Daten bewegen zu können.
Werden keine Körperbewegungen simuliert, sondern beispielsweise Gesichtszüge, kann wie [11] zeigt durch eine geeigneten Repräsentation in Form von Linearkombinationen aus Basisgesichtszügen das Mesh direkt verwendet werden.

2.4 Aufbereitung der Daten

Die Aufbereitung der Mocap-Daten ist ein wichtiger Schritt vor der Vereinigung mit dem menschlichen Modell, das bewegt werden soll. Dieser Schritt stellt sicher, dass einige häufige Probleme bereits vorzeitig vermieden werden können. Ein Grundproblem stellen Rauschen und fehlende Werte in den aufgenommenen Daten dar [4, 30]. Artefakte und Störungen können beispielsweise bei optischen Systemen durch die Verdeckung eines Sensors, reflektierende Fläche, fehlerhaftes Zuweisen von Markern zu Positionen im Skelett oder Bewegungsunschärfe auftreten [4]. Neben der zeit- und kostenintensiven Neuaufnahme kann geringes Rauschen durch Tiefpassfilter [4] oder den Einsatz eines Kalman-Filters [30] ausgeglichen werden.
Neben dieser Erstkorrektur, die je nach Daten nicht immer nötig sein muss, stellt die Aufbereitung sicher, dass die Daten den Erwartungen entsprechen. Beispielsweise kann in den Rohdaten ein Schweben des Skeletts auftreten, sodass zwischen Boden und Füßen ungewollter Leerraum entsteht. Neben der Korrektur dieses Schwebens, kann zusätzlich je nach Anwendungsfall festgelegt werden, ob das Skelett als Ganzes eine Bewegung besitzt oder für spätere manuelle Anpassungen auf einer Stelle verweilend seine Knochenbewegungen ausführt. Außerdem kann festgelegt werden, ob die Animation eine Schleife durchlaufen soll und dementsprechend, beispielsweise mittels Autokorrelation, angepasst werden. Für diese Aufgaben hilfreich ist ein Tool namens BVHacker[10].

2.5 Retargeting

Unter *Retargeting* wird die Übertragung einer animierten Bewegung von einem Skelett oder Charakter auf einen anderen bezeichnet [14], wobei die Natürlichkeit erhalten bleiben soll. Dies ist der Schritt, indem die Mocap-Daten mit dem Menschmodell verbunden werden. Retargeting ist nur dann nicht nötig, wenn beide Skelette dieselben sind, was durch Vorgaben der Systeme nur selten der Fall ist. Durch die unterschiedliche geometrische Ausprägung verschiedener Menschmodelle und derer Skelette muss die Animation angepasst werden, um natürlich zu wirken. Außerdem kann die hierarchische Struktur der beiden Skelette verschieden sein, sodass eine Abbildung von einem Gelenk auf ein möglichst Ähnliches im zweiten Skelett erfolgen muss. Beide Aufgaben zusammen wurden ohne Keyframing erstmals von Monzani et al. [27] gelöst, indem eine Zwischenrepräsentation genutzt wird. Dieses Zwischenskelett hat die gleiche hierarchische Struktur wie das Zielskelett des Menschmodells, übernimmt jedoch die Orientierungen des Mocap-Skeletts. Die Zuordnung von Knochen des Ausgangs- zum Zielskelett muss jedoch manuell erfolgen. Ebenso müssen mögliche Einschränkungen für bestimmte Gelenke und Knochen manuell definiert werden, um beispielsweise festzulegen, dass eine Hand immer ein Objekt berühren muss. Diese Einschränkungen werden anschließend mithilfe von Inverser Kinematik berücksichtigt. Unter Inverser Kinematik wird die Anpassung von Gelenkwinkelstellungen verstanden, wobei das Ziel ist, gegebenen kartesischen Zielkoordinaten für ein bestimmtes Gelenk möglichst nahe zu kommen [10]. Durch den Einsatz inverser Kinematik, um alle Einschränkungen zu erfüllen, kann es jedoch laut [24] zur Veränderung der Charakteristik der Originalbewegung kommen. Deshalb wird eine Verbesserung vorgeschlagen, die vorab die kinematisch korrekten Einschränkungen von bestimmten Gelenkpositionen aus dem Mocap-Skelett berechnet und so Probleme verhindern kann. Die auf Monzani basierenden Verfahren finden vielfach Anwendung, sowohl in anderen wissenschaftlichen Arbeiten [7], als auch in Anwendungen für Animateure wie den Blender Motion Capture Tools[11] [9] oder in kommerziellen Anwendungen [17].
Die notwendigen manuellen Eingriffe, insbesondere das Erstellen einer Abbildung der Gelenke oder Knochen zwischen den Skeletten, sind kaum vermeidbar und aus diesem Grund auch in kommerziellen Anwendungen, wie dem Autodesk MotionBuilder[12] notwendig. Ansätze, diese Eingriffe zu minimieren basieren auf Heuristiken. Feng et al. [12] nutzen Ähnlichkeiten im hierarchischen Aufbau, der Symmetrie und eine Schlüsselwortsuche, um eine automatische Korrespondenz zwischen den Skeletten zu erstellen. Somit kann oftmals auf ein manuelles Eingreifen verzichtet werden beziehungsweise es müssen lediglich einzelne Korrekturen erfolgen.

2.6 Probleme und Schwierigkeiten

Während der Übertragung von Mocap-Daten auf virtuelle Menschmodelle können ver-

[6] http://resources.mpi-inf.mpg.de/HDM05/, 10.09.2015
[7] https://www.mixamo.com/3d-animations, 10.09.2015
[8] https://secure.axyz-design.com/metropoly-3d-people, 10.09.2015
[9] http://www.makehuman.org/, 12.09.2015
[10] http://www.bvhacker.com/, 15.09.2015
[11] http://wiki.blender.org/index.php/Extensions:2.6/Py/Scripts/Animation/Motion_Capture_Tools, 20.09.2015
[12] http://www.autodesk.com/products/motionbuilder/overview, 20.09.2015

schieden herausfordernde Probleme auftreten. Im einfachsten Fall sind die Koordinatenachsen der beiden Skelette nicht kompatibel. Dies kann durch das korrekte Abstimmen der Achsen korrigiert werden. Ebenso können Probleme durch eine Fehlzuordnung der Gelenke aufreten. Hier kann nur ein erneutes Testen mit einer anderen Konfiguration Abhilfe schaffen. Bei weitem das häufigste Problem ist jedoch das sogenannte *Foot Skating* bei dem die Füße bei Gehbewegungen über den Boden rutschen. Dies hat zur Folge, dass der Charakter statt bei jedem Schritt einen festen Standpunkt zu haben eher den Eindruck erweckt Schlittschuh zu fahren. Dieser Effekt tritt nach [23] beispielsweise durch schlecht kalibrierte Aufnahmegeräte, fehlerhafte Skelettabbildungen oder durch das bereits beschriebene Retargeting auf. Insbesondere im letzten Fall tritt das Problem auf, wenn die Eigenbewegung des Mocap-Skeletts auf ein kürzeres Zielskelett übertragen wird, da so bei gleicher Schrittanzahl und kürzeren Schrittlängen die gleiche Strecke zurückgelegt werden muss [24]. Die Allgemeine Lösung ist, die Fußbewegung einzuschränken, indem diese fixiert und durch Inverse Kinematik erhalten wird oder die Eigenbewegung des Skeletts anzupassen. Durch ersteren Ansatz müssen oft harte Gelenkwinkelanpassungen durchgeführt werden, weshalb [23, 19] die Tatsache nutzen, dass Größenänderungen von unter 20 Prozent im Normalfall nicht auffallen [19]. Durch die langsame und eingeschränkte Verlängerung und Verkürzung der Bein- und Fußknochen ist es so möglich eine unauffälligere Korrektur des Foot Skatings zu erreichen. Dieser Ansatz ist jedoch nicht immer anwendbar, da teilweise fixe Knochenlängen in den verarbeitenden Anwendungen nötig sind. In einem solchen Fall kann nach dem zweiten Ansatz aus den fixen Standpunkten in den Mocap-Daten die dazu passende Eigenbewegung berechnet werden. Dabei ändert sich die zurückgelegte Strecke je nach Größe des Rigs. Die größte Schwierigkeit bei der Übertragung von Animationen zwischen verschiedenen Charakteren ist den Realismus zu erhalten. Beispielsweise unterscheiden sich der Gang eines weiblichen und eines männlichen Schauspielers teils deutlich. Soll nun eine Gang-Animation auf beide Geschlechter angewendet werden können, sind meist Anpassungen nötig. Eine Möglichkeit solche Änderungen in Mocap-Animationen einfließen zu lassen, ist das sogenannte Motion Warping [34]. Hierbei wird mittels eines Keyframes eine Randbedingung gesetzt. Durch das Addieren eines Offsets auf die Bewegungskurven der Gelenke wird diese Bedingung erfüllt. Dabei muss die Dauer der Wirksamkeit dieser Randbedingung angegeben werden.

3 Analyse von Simulationsumgebungen

Für die Umsetzung einer Simulationsumgebung, die die Aspekte des vorigen Abschnitts berücksichtigt, müssen zunächst die bestehenden Plattformen betrachtet und analysiert werden. Diese könnten als Basis für eine Umsetzung dienen. Grundsätzlich sind zwei Plattformtypen zu unterscheiden, die durch die jeweiligen Eigenschaften als Basis geeignet scheinen. Bevor näher auf diese eingegangen wird, werden im folgenden Abschnitt Kriterien definiert, die eine Simulationsanwendung erfüllen sollte, um menschliche Bewegungen für die Entwicklung von kognitiv technischen Komponenten der Bewegungswahrnehmung simulieren zu können.

3.1 Kriterien

Um die Realisierung einer Simulationsumgebung zu ermöglichen, sind einige Kriterien zwingend zu erfüllen, die im folgenden hervorgehoben dargestellt werden. Das System muss **gängige Datenformate für Animationen laden** und in einer 3D-Umgebung abspielen können. Nur so ist eine Nutzung von Mocap-Daten gewährleistet. Von Vorteil könnte außerdem eine direkte Live-Übertragung von Bewegungen sein. Auch das **Importieren von Menschmodellen** ist eine grundlegende Voraussetzung für eine flexible Simulationslösung.

Die Aufbereitungsalgorithmen für Mocap-Daten müssen dagegen nicht zwingend integriert werden, da an dieser Stelle spezialisierte Anwendungen mehr Möglichkeiten bieten. Wünschenswert ist dagegen eine **integrierte Retargeting-Lösung**, sodass Menschmodell und Animation beliebig und unabhängig ausgetauscht werden können. Für die Anwendung im Bereich Computer Vision sind schließlich noch einige weitere Kriterien zu erfüllen. Voraussetzung ist die Möglichkeit, **Referenzdaten für das aktuelle Bild auslesen** zu können. Diese sollen dabei flexibel definiert werden können, um neben Bounding Boxen beispielsweise auch Gelenkpositionen, Geschwindigkeiten und Objektzugehörigkeiten registrieren zu können. Eng mit dieser Anforderung verbunden ist die Abbildung von Sensoren und kinematisch korrekten Fahrzeugen, die insbesondere im Automobilbereich eine umfassendere Sensorsimulation ermöglichen würde. In einem solchen Anwendungsfall wären auch manuelle oder algorithmische Steuerungsmöglichkeiten für Fahrzeuge sinnvoll, damit Feedback Loops realisiert werden können. Essentiell für die Analyse von Bildverarbeitungsalgorithmen ist die Möglichkeit, **Video- oder Bildaufnahmen** des Simulationsszenarios zu erstellen. Eine **realitätsnahe Darstellung** kann grundsätzlich von Vorteil sein [33]. Dies gilt einerseits für die grafische Darstellung und andererseits für die Berechnung physikalisch korrekter Ereignisse. Letzteres setzt eine **Physikengine** voraus. Besonders die Darstellung verschiedenster Umgebungsbedingungen ist ein Vorteil des Simulationsansatzes. Aus diesem Grund sollten **Wetteränderungen, Kamerafilter und ein geeigneter Szenario-Editior** eingesetzt werden können. Abschließend sind einige Kriterien wünschenswert um die Anwendungsgebiete der Simulationsumgebung zu erhöhen. Dazu gehören die zeitgleiche Abbildung mehrere Kameras, das Arbeiten über Netzwerkprotokolle, sowie für beliebige Anpassungen das Vorhandensein einer Erweiterungsmöglichkeit mittels verbreiteten Programmiersprachen. Die Einsatzfähigkeit würde außerdem durch einen geringen Einarbeitungsaufwand positiv beeinflusst.

3.2 Bestehende Simulationsumgebungen

Unter diesem Typ werden zumeist kommerzielle Produkte zusammengefasst, die das explizite Ziel haben, eine physikalisch korrekte Simulation zu erzeugen. Diese Anwendungen sind für einen speziellen Anwendungsfall entwickelt, sodass Lösungen für Ergonomieuntersuchungen (Jack, 3D SSPP, Santos Human) oder Simulationen für die Entwicklung von Sensor-gesteuerten Fahrzeugen (Pro-Sivic, PreScan, CarSim) existieren. Sie bieten exzellente Möglichkeiten im Bereich der Referenzdatengenerierung, der Videoaufnahmen einzelner Kameraperspektiven, der flexiblen Szenariogestaltung und der Abbildung von speziellen Sensoren. Problematischer ist der Import neuer Animationen. Dies ist beispielsweise bei 3D SSPP, CarSim, Pro-Sivic und bislang auch PreScan nicht vorgesehen und durch die kommerzielle Ausrichtung nicht integrierbar. Diese Produkte sind für die Erfüllung der nötigen Anforderungen nicht geeignet. Im Bereich Fahrzeugsensorsimulation scheint von den getesteten Produkten PreScan die besten Möglichkeiten zu bieten, da es mit DAE bereits ein gängiges Format für 3D-Modelle unterstützt. In Zukunft könnten auch im Format integrierbare Animationen beim Importieren unterstützt werden.

3.3 Spiel-Engines

Spiel-Engines sind für die Entwicklung von Computerspielen gedachte Frameworks, die grundlegende Funktionen bereitstellen. Weit verbreitet sind insbesondere Unity, die Unreal Engine und die CryEngine. Da für moderne Computerspiele eine realitätsnahe grafische Darstellung und eine flexible Anpassung von Szenarios nötig ist, können Spiel-Engines beinahe alle Kriterien aus Abschnitt 3.1 erfüllen. In erster Linie ist das Importieren und Verarbeiten von Animationen der gängigen Dateiformate meist essentiel-

ler Bestandteil. Retargeting-Lösungen sind bereits integriert, sodass Animationen innerhalb der Engines von verschiedenen Rigs genutzt werden können. Das Generieren von Referenzdaten und Sensordaten, sowie die Aufnahme von Kamerabildern ist dagegen nicht integriert, da dies für Computerspiele nicht benötigt wird. Auch ein physikalisch korrektes Fahrzeugmodell ist standardmäßig nicht enthalten. Da alle Engines im Vergleich zu kommerziellen Simulationsumgebungen erweiterbar und programmierbar sind, könnten diese Funktion variabel hinzugefügt werden. Von den drei genannten Engines bietet Unity die meisten Möglichkeiten im Bereich Retargeting und erlaubt eine Multikameraansicht.

4 Anwendungen in der Bildverarbeitung

Nachdem die vorigen Abschnitte verdeutlichen, welcher Aufwand für die realistische Simulation von menschlichen Bewegungen erforderlich ist, wird nun auf den dadurch entstehenden Nutzen im Bereich der Bewegungswahrnehmung und damit der Bildverarbeitung eingegangen. Dieses Gebiet ist nur eines, indem solche Simulationen von Vorteil sind. Die Einleitung deutet bereits einige Aspekte an. Prinzipiell können die Vorteile für die Bildverarbeitung in zwei Zielgebiete aufgeteilt werden, die in den beiden folgenden Abschnitten betrachtet werden. Zunächst kann mittels menschlichen Bewegungssimulationen eine Analyse beziehungsweise ein Vergleich von Algorithmen stattfinden. Dies führt auf lange Sicht zu deren Optimierung. Die Bewegungsdaten und -bilder können jedoch viel direkter für eine Optimierung eingesetzt werden. Dies stellt das zweite Gebiet dar, in welchem aktuell Forschung betrieben wird.

4.1 Analyse und Vergleich

Die Analyse und der Vergleich von Algorithmen sind ein wichtiges Mittel in der Forschung, um Fortschritte nachzuweisen und Verbesserungen zu erreichen. So werden im Bereich der Bildverarbeitung viele Datensätze und passende Bewertungskriterien bereitgestellt, um gezielt eine Aufgabe, wie beispielsweise die Fußgängererkennung aus fahrenden Autos, bewerten zu können. Die Erstellung eines solchen Datensatzes ist jedoch zeitintensiv, da manuelle Annotationen beispielsweise in Form von Bounding Boxes erfolgen müssen. Nur so kann die Aussage des Algorithmus mit der wahren Begebenheit verglichen werden. Ein Ansatz, den Aufwand gering zu halten, ist der Einsatz von Online-Tools, die es erlauben die manuelle Arbeit, teils gegen geringe Bezahlung, von Anderen durchführen zu lassen [3]. Doch obwohl auf dieses Weise viele annotierte Daten verfügbar werden, ist keine Sicherheit vorhanden, ob diese Daten qualitativ hochwertig sind [3, 33]. Simulationen scheinen hier die Lösung zu sein. Sie erlauben eine gezielte Generierung des Bildes und synchroner und präziser Referenzdaten, da die Daten für die Bilderstellung zwangsläufig benötigt werden. Durch die Möglichkeit einfacher Änderungen im Renderingablauf kann eine hohe Variabilität und Repräsentativität der Daten erzeugt werden. So ist es möglich Szenenhintergründe zu verändern, Objekte anzupassen oder auszutauschen und die erzeugten Referenzdaten zu verfeinern. Außerdem können gezielt Abläufe getestet werden, um bestimmte Bedingungen abzudecken [33]. Auf diese Weise lassen sich zum Beispiel im Straßenverkehr verbotene oder gefährliche Situationen testen. Zusätzlich ist für erste Tests die Anschaffung teurer Hardware nicht nötig. Obwohl realitätsnahe virtuelle Daten nicht genau die Realität widerspiegeln, zeigen Anwendungen in der Segmentierung [18, 31], der Posen- und Aktionserkennung [29], bei der Detektion [16, 8], im Tracking [32, 2] und der Analyse des optischen Flusses [26, 1], dass der Einsatz sinnvoll ist. All diese Vorteile führen dazu, dass durch die Analyse eines Algorithmus mit virtuellen Daten gezielte Verbesserungen erreicht werden können, indem spezielle Fälle berücksichtigt werden. Darüber hinaus ist ein weniger aufwendigerer, jedoch anspruchsvollerer Vergleich zwischen Algorithmen möglich. Insbesondere einzelne Körper- oder Gelenkstellungen automatisch annotieren zu können ist im Bereich der menschlichen Bewegungen ein enormer Vorteil im Gegensatz zur kaum möglichen Annotation solcher Gegebenheiten per Hand.

4.2 Optimierung

Im vorigen Abschnitt wurden insbesondere die Flexibilität und die Einfachheit in der Generierung großer Datenmengen mit Referenzdaten betrachtet, um Algorithmen zu evaluieren und zu analysieren. Dabei wird davon ausgegangen, dass die meist überwachten Lernverfahren zu diesem Zweck in derselben (virtuellen) Domäne trainiert und getestet werden. Bei der direkten Optimierung wird ein Schritt weiter gegangen, indem Simulationsdaten genutzt werden, um einen Algorithmus für reale Umgebungen zu trainieren. Der große Vorteil ist die enorme Anzahl an Daten, die so für das Training genutzt werden kann, da viele Lernverfahren mit einer steigenden Zahl variabler Trainingsdaten besser abschneiden. Eines der bekanntesten Verfahren, das sich dies zu Nutze macht, ist die Microsoft Kinect, die mit einer enormen Menge an menschlichen Posen von synthetisch generierten Bildern trainiert wurde [13]. Ähnliche Verfahren werden in [7, 31] beschrieben, die genauer auf die Datengenerierung eingehen. In diesem Fall sind die Daten lediglich Tiefenbilder, die recht einfach generiert werden können, das Prinzip kann jedoch allgemeiner genutzt werden. So wurde beispielsweise eine gute Leistungsfähigkeit für eine nach diesem Schema trainierte Fußgängererkennung nachgewiesen [33]. Zu Beachten ist das Auftreten des sogenannten *Dataset Shift Problems*. Dieses ist gegeben, wenn der Trainingsdatensatz nicht zum Kontext oder Anwendungsfall während des Testlaufes passt [33]. Bei der Nutzung von virtuellen Daten in einem realen Kontext ist dies eindeutig, das Problem tritt jedoch auch auf, wenn ein System mit einer Kamerakonfiguration trainiert und mit einer anderen angewendet wird, wobei in beiden Fällen Realdaten zum Einsatz kommen. Die Lösung dieses Problems wird allgemein als *Domain Adaption* bezeichnet, wobei zur Umsetzung verschiedene Methoden genutzt werden. Einerseits ist das Ändern oder Anpassen der Merkmale an die neue Domäne möglich. Andererseits werden häufig die genutzten Lernverfahren insbesondere Support Vector Machines angepasst, um die Domain Adaption durchzuführen [6]. In [33] werden unterschiedliche Verfahren vorgestellt, die teils überwacht und teils unüberwacht arbeiten. Als gute Kombination hat sich erwiesen, das auf virtuellen Bildern gelernte Modell auf reale Bilder anzuwenden, dabei jedoch einen Bereich der Unsicherheit bei der Klassifizierung zu definieren [33]. Testdaten, die in diesen Bereich fallen, werden anschließend manuell annotiert und für ein spezifischeres Training des Modells mit einem augmentierten Merkmalsraum genutzt. Das vollständige Umgehen manueller Annotationen ist beispielsweise mit einem transductive SVM (T-SVM) [22] möglich, wobei die Trainingszeit jedoch erheblich ansteigt und schwierig zu bestimmende Parameter geschätzt werden müssen [33]. Eine Alternative scheint die Methode nach [15] zu sein, die die sogenannten *Geodesic Flows* nutzt, um interpolierte Zwischenräume der Quell- und Zielmerkmalsräume zu generieren. Eine detailliertere Übersicht zu weiteren Methoden findet sich in [28].

Neben der Optimierung eines Algorithmus durch effizienteres Lernen mit mehr Trainingsdaten, können weitere Parameter eines Systems optimiert werden, indem eine Simulation genutzt wird. Beispielsweise ist es in einer solchen Umgebung auf einfache Weise möglich dieselbe Szene mehrfach zu simulieren, jedoch die Kamerapositionen oder deren Einstellungen zu ändern. Dies erlaubt eine Optimierung dieser Randbedingungen vor einer physikalischen Installation. So kann die Sichtbarkeit von einzelnen, wichtigen Merkmalen verbessert werden [31].

5 Fazit

Die Simulation menschlicher Bewegungen erlaubt eine verbesserte und flexiblere Ent-

wicklung von kognitiv technischen Komponenten der Bewegungswahrnehmung. Die gezielte Analyse spezifischer, real aufwendiger Szenarios und die einfache Anpassung von Umgebungen und Bedingungen in der Simulation können zu präzisen Verbesserungen führen. Die direkte Optimierung durch große simulierte Datenmengen und Domain Adaption im Bereich der überwachten Lernverfahren birgt erhebliches Potential zur Effizienz- und Leistungssteigerung. Trotz dieser positiven Auswirkungen, ist das nötige Vorgehen, von Mocap-Daten zu einer Simulation zu gelangen, aufwendig und teils nur manuell zu bewältigen. Darüber hinaus ist speziell für die Simulation menschlicher Bewegungen bei gleichzeitiger Generierung von Referenzdaten keine bestehende Simulationsumgebung ohne weiteres einsetzbar. Weitere Arbeiten in diesem Bereich könnten diese Lücke schließen und auf diese Weise einen möglichst einfachen Simulations- und Optimierungsprozess ermöglichen.

Literatur

[1] E. L. Andrade und R. B. Fisher. Simulation of crowd problems for computer vision. In *First International Workshop on Crowd Simulation*, Bd. 3, S. 71–80, 2005.

[2] P. Baiget, X. Roca, und J. Gonzàlez. Autonomous virtual agents for performance evaluation of tracking algorithms. In *Articulated Motion and Deformable Objects*, S. 299–308. Springer, 2008.

[3] T. L. Berg, A. Sorokin, G. Wang, D. A. Forsyth, D. Hoiem, I. Endres, und A. Farhadi. It's all about the data. *Proceedings of the IEEE*, 98(8):1434–1452, 2010.

[4] B. Bradwell und B. Li. A tutorial on motion capture driven character animation. In *Proceedings of the Eighth IASTED International Conference*, Bd. 630, S. 356, 2008.

[5] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] P. P. Busto, J. Liebelt, und J. Gall. Adaptation of synthetic data for coarse-to-fine viewpoint refinement. *British Machine Vision Conference (BMVC'15)*, 2015.

[7] K. Buys, J. Hauquier, C. Cagniart, T. Tuytelaars, und J. De Schutter. Virtual data generation based on a human model for machine learning applications. In *Proceedings of the international digital human modeling conference*, 2013.

[8] C. Chen, A. Seff, A. Kornhauser, und J. Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. *arXiv preprint arXiv:1505.00256*, 2015.

[9] B. Cook. Improving motion capture import & workflow, http://wiki.blender.org/index.php/User:Benjycook/GSOC/Proposal. 2011, Zugriff: 10.09.2015.

[10] P. Corke. *Robotics, vision and control: fundamental algorithms in MATLAB*, Bd. 73. Springer Science & Business Media, 2011.

[11] C. Curio, M. Breidt, M. Kleiner, Q. C. Vuong, M. A. Giese, und H. H. Bülthoff. Semantic 3d motion retargeting for facial animation. In *Proceedings of the 3rd symposium on Applied perception in graphics and visualization*, S. 77–84. ACM, 2006.

[12] A. Feng, Y. Huang, Y. Xu, und A. Shapiro. Fast, automatic character animation pipelines. *Computer Animation and Virtual Worlds*, 25(1):3–16, 2014.

[13] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, und A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *2011 IEEE International Conference on Computer Vision (ICCV)*, S. 415–422. IEEE, 2011.

[14] M. Gleicher. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, S. 33–42. ACM, 1998.

[15] R. Gopalan, R. Li, und R. Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(11):2288–2302, 2014.

[16] D. Gruyer, S. Pechberti, und S. Glaser. Development of full speed range acc with sivic, a virtual platform for adas prototyping, test and evaluation. In *Intelligent Vehicles Symposium (IV), 2013 IEEE*, S. 100–105. IEEE, 2013.

[17] S. Guo, R. Southern, J. Chang, D. Greer, und J. J. Zhang. Adaptive motion synthesis for virtual characters: a survey. *The Visual Computer*, 31(5):497–512, 2015.

[18] V. Haltakov, C. Unger, und S. Ilic. Framework for generation of synthetic ground truth data for driver assistance applications. In *Pattern Recognition*, S. 323–332. Springer, 2013.

[19] J. Harrison, R. A. Rensink, und M. Van De Panne. Obscuring length changes during animated motion. In *ACM Transactions on Graphics (TOG)*, Bd. 23, S. 569–573. ACM, 2004.

[20] N. Hasler. Motion capture. In K. Ikeuchi, editor, *Computer Vision*, S. 495–498. Springer US, 2014.

[21] C. Henssler. Techniken zur deformation von virtuellen menschmodellen. In *Informatics Inside 2015, Tagungsband*, S. 8–15. Hochschule Reutlingen, 2015.

[22] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, Bd. 99, S. 200–209, 1999.

[23] L. Kovar, J. Schreiner, und M. Gleicher. Footskate cleanup for motion capture editing. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation*, S. 97–104. ACM, 2002.

[24] W. Lu, Y. Liu, J. Sun, und L. Sun. A motion retargeting method for topologically different characters. In *Computer Graphics, Imaging and Visualization, 2009. CGIV'09. Sixth International Conference on*, S. 96–100. IEEE, 2009.

[25] M. M. S. Maddock. Motion capture file formats explained. *Department of Computer Science, University of Sheffield*, 2001.

[26] S. Meister und D. Kondermann. Real versus realistically rendered scenes for optical flow evaluation. In *Electronic Media Technology (CEMT), 2011 14th ITG Conference on*, S. 1–6. IEEE, 2011.

[27] J.-S. Monzani, P. Baerlocher, R. Boulic, und D. Thalmann. Using an intermediate skeleton and inverse kinematics for motion retargeting. In *Computer Graphics Forum*, Bd. 19, S. 11–19. Wiley Online Library, 2000.

[28] V. M. Patel, R. Gopalan, R. Li, und R. Chellappa. Visual domain adaptation: A survey of recent advances. *Signal Processing Magazine, IEEE*, 32(3):53–69, 2015.

[29] C. Schlette, A. G. Buch, E. E. Aksoy, T. Steil, J. Papon, T. R. Savarimuthu, F. Wörgötter, N. Krüger, und J. Roßmann. A new benchmark for pose estimation with ground truth from virtual reality. *Production Engineering*, 8(6):745–754, 2014.

[30] H. J. Shin, J. Lee, S. Y. Shin, und M. Gleicher. Computer puppetry: An importance-based approach. *ACM Transactions on Graphics (TOG)*, 20(2):67–94, 2001.

[31] Stephan Irgenfried and Frank Dittrich and Heinz Wörn. Realization and evaluation of image processing tasks based on synthetic sensor data: 2 use cases. In M. H. Puente León, Fernando, editor, *Forum Bildverarbeitung 2014*, S. 11, 2014.

[32] G. R. Taylor, A. J. Chosak, und P. C. Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, S. 1–8. IEEE, 2007.

[33] D. Vázquez Bermúdez, A. M. López Peña, D. Ponsa Mussarra, et al. Domain adaptation of virtual and real worlds for pedestrian detection. 2013.

[34] A. Witkin und Z. Popovic. Motion warping. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, S. 105–108. ACM, 1995.

Eye-Tracking-Studie zu zwei E-Learning-Materialien mit Fokus auf die Aufmerksamkeitssteuerung

Veronika Rein
Reutlingen University
Veronika.Rein@Student.
Reutlingen-University.DE

Abstract

Die Erforschung der Informationsverarbeitung findet in der Industrie, der Forschung sowie der Medienpsychologie ihre Anwendung [23, S. 25]. Forschungsaspekte der Medienpsychologie sind die Analyse der visuellen Informationsaufnahme und der Medienrezeption. Diese Aspekte werden in der wissenschaftlichen Vertiefung anhand von zwei E-Learning-Materialien in einer Eye-Tracking-Studie untersucht. Die Resultate der Studie sollen aufzeigen, inwieweit sich die Aufmerksamkeit der Rezipienten von der Darbietung der Informationen steuern lässt. Die Untersuchungsresultate werden anhand von Tracking-Daten ausgewertet. Zuletzt findet eine kritische Diskussion zur Untersuchung statt. Darauf folgt das Fazit.

Schlüsselwörter

Eye-Tracking, Aufmerksamkeitssteuerung, Informationsverarbeitung

CR-Kategorien

[I.] Computing Methodologies; [I.4] Image Processing and Computer Vision; [I.4.8] Scene Analysis: Tracking

---

Betreuer Hochschule: Prof. Dr. rer. nat. Dopatka
Hochschule Reutlingen
Frank.Dopatka@Reutlingen-University.de

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

1 Einleitung

Aufmerksamkeit lässt sich in drei Kategorien unterteilen: selektive, fokussierte und geteilte. [9, S. 33] Bei der selektiven Aufmerksamkeit werden unbewusst selektierte Umgebungsaspekte wahrgenommen. Die geteilte Aufmerksamkeit wird gleichzeitig auf verschiedene Aspekte verteilt. [9, S. 34] Bei der fokussierten Aufmerksamkeit werden die Umgebungsaspekte bewusst wahrgenommen. Die Aufmerksamkeit ist bei der Informationsaufnahme entscheidend. Sie ist auch für die Haftung der Informationen im Gehirn mitverantwortlich. [4, S.11] Bei einem durchdachten Einsatz und der Steuerung der fokussierten Aufmerksamkeit können Informationen so präsentiert werden, dass sie nachhaltig im Gedächtnis bleiben. [9, S.9]

In dieser wissenschaftlichen Arbeit wird untersucht, in welcher Reihenfolge visuelle Informationen aufgenommen und verarbeitet werden. Das Ziel dabei ist es, festzustellen, ob sich der Blickverlauf bei der Informationsaufnahme durch stimulusspezifische Eigenschaften wie unterschiedliche Beschaffenheit der Lernmaterialien steuern lässt. Es lassen sich folgende Forschungsfragen ableiten:

1) Richtet sich der Blick der Probanden bei einer dynamischen Darstellung unverzüglich auf ein Element, sobald dieses im Gesamtbild erscheint?

2) Werden in einer statischen Darstellung alle Elemente von den Probanden fixiert?

Letztendlich ist der Zusammenhang der Informationsaufnahme mit den Ergebnissen einer Abfrage zu den Inhalten zu erforschen. Angenommen wird dabei, dass sich der Blickverlauf bei einer dynamischen Darstellung der Information besser steuern lässt als bei einer statischen Darstellung. Um diesen Sachverhalt zu erforschen, bietet sich Eye-Tracking an. Als Instrument für die Blickverfolgung ermöglicht es, die Aufmerksamkeit und den Fluss der Informationsaufnahme der Rezipienten zu untersuchen.

2 Informationsaufnahme

Bei der Informationsaufnahme gelangt ein Reiz in das Gehirn. Dort wird er weiterverarbeitet. Wenn es sich um eine visuelle Informationsaufnahme handelt, dann dient das Blickverhalten als Grundlage dafür. Dieses erfolgt zwar reizgesteuert, kann aber durch stimulusspezifische Eigenschaften kontrolliert werden. [11, S.161]

2.1 Visuelle Wahrnehmung

„Am Anfang des visuellen Wahrnehmungsprozesses, noch vor der sensorischen Registrierung und Verarbeitung über das Auge, steht die visuelle Aufmerksamkeit.“ [11, S.164] Die durch visuelle Aufmerksamkeit gesteuerte Blickkontrolle der Wahrnehmung ist die zentrale Größe für die Messung des Blickverlaufs über Eye-Tracking. [11, S.164f]

Die menschliche Wahrnehmung ist ein Aufnahmeprozess von Sinnesreizen. Visuelle Reize werden aufgenommen, indem Licht in bestimmten Wellenlängen durch die Linse des Auges fällt, dort gebrochen wird und auf der Netzhaut ein Bild hinterlässt. [13, S. 73] Wahrnehmung funktioniert, abhängig von der Aufgabe, nach zwei Prinzipien: entweder Bottom-Up- oder Top-Down. Die Bottom-Up-Verarbeitung entspricht den ersten Fixationen bei der Bildbetrachtung. Sie ist reizgesteuert und kann vom Menschen nicht beeinflusst werden. Die Top-Down-Verarbeitung ist aufgabenbasiert und kann vom Menschen beeinflusst werden. Sie hängt von kognitiven Prozessen ab, die erst nach einigen Sekunden bei der Betrachtung hinzukommen. [6, S.4]

2.1.1 Augenbewegungen

Augenbewegungen lassen sich in Sakkaden und Fixationen einteilen. Fixationen sind stabilisierende Bewegungen, die mindestens 0,2 Sekunden dauern. Es handelt sich dabei um einen Stillstand des Blickes mit einer fokussierten Aufmerksamkeit auf ein Wahrnehmungsobjekt. Während dem Stillstand nimmt das menschliche Gehirn Informationen auf. [3, S.279], [1, S.146] Ausschließlich fixierte Informationen können verarbeitet werden. [13, S. 71] Dabei wird die Dauer einer Fixation als Maß für die Dauer der Informationsverarbeitung interpretiert. [7, S. 156] McConkie et al. (1985) haben für das Lesen gezeigt, dass die Verarbeitungszeit kürzer als die Fixationsdauer sein kann. [22]

Sakkaden sind Sprünge von einer Fixation zur nächsten und dauern wenige Millisekunden. Innerhalb dieser kurzen Zeit erfolgt vom Gehirn keine Wahrnehmung. [1, S.146] Wenn die Erforschung eines Bildes ohne ein bestimmtes Ziel verläuft, handelt es sich um explorative Sakkaden. Dagegen werden Reflexsakkaden auf neu erscheinende sensorische Stimuli hin generiert. [19, S. 12] Die Augen bewegen sich in Richtung dieser Stimuli.

2.2 Kontrolle visueller Wahrnehmung

Der zentrale Prädikator des Sakkaden-Fixations-Verlaufs ist die visuelle Aufmerksamkeit. Diese kann durch die Charakteristika eines Stimulus gelenkt werden. [21, S.171] „So wird ein plötzlich im Wahrnehmungsraum auftauchendes Objekt nahezu automatisch visuelle Aufmerksamkeit und damit Blickfokussierung auf sich ziehen […].“ [21, S.171] Dafür ist der Orientierungsreflex verantwortlich. Die externe Blicklenkung wird als exogene Kontrolle bezeichnet (auch stimulus-driven nach Godijn und Theeuwes, 2003). [21, S.172] Der exogenen Kontrolle steht die endogene Kontrolle gegenüber (auch goal-driven nach Godijn und Theeuwes, 2003). Sie basiert auf konkreten Wahrnehmungszielen und bezieht die Intention einer visuellen Suche ein. [12, S.12]

3 Eye-Tracking

Eye-Tracking ist eine apparative Erhebungsmethode zum Aufzeichnen der aus Fixationen und Sakkaden bestehenden Blickbewegungen eines Menschen. [1, S.1] Der Blickverlauf eines Rezipienten wird vom Eye-Tracker beim Betrachten eines Wahrnehmungsobjekts registriert und festgehalten. Dabei werden die Sakkaden ebenso festgehalten wie die Dauer der Fixationen. [2, S.6] Nach der Erfassung der Blickbewegungen ermöglicht Eye-Tracking, diese qualitativ und quantitativ zu beschreiben und zu analysieren. [11, S.161] Nach der Eye-Mind-Hypothese liefert Eye-Tracking Erkenntnisse über den Prozess der Medienrezeption und ermöglicht Rückschlüsse auf die Bewältigung kognitiver Aufgaben. [23]

Die Abbildung 1 zeigt die Analysemöglichkeiten mithilfe der Blickregistrierung und -Verfolgung im Bereich des E-Learnings.

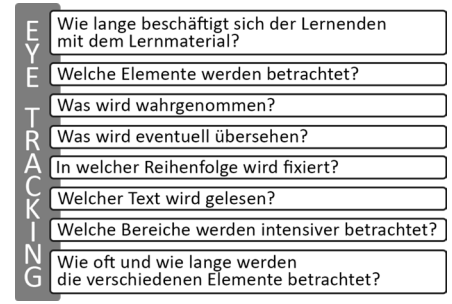

Abbildung 1: Analysemöglichkeiten des Eye-Trackings [vgl. 1, S.151]

3.1 Videobasierte Eye-Tracker

Zu den videobasierten Eye-Trackern zählen der invasive, mobile Headmounted-Eye-Tracker und der passive, stationäre Tablemounted-Eye-Tracker. Der Headmounted-Eye-Tracker wird am Körper getragen. Dabei können die Kameras zur Blickregistrierung in einen Helm oder in eine Brille integriert werden. [11, S.180] Bei dem Tablemounted-Eye-Tracker ist entweder die Blickaufzeichnungshardware in den Monitor integriert oder der Eye-Tracker unter dem Bildschirm angebracht. Der Eye-Tracker arbeitet berührungslos und misst videobasiert den Cornealen Reflex auf der Hornhaut des Auges (Abb. 2). [11, S.178] Diese Technik erlaubt eine hohe Bewegungsfreiheit, wobei der Eye-Tracker den Fixationsort nicht verliert.

Hinsichtlich der Auswertung der erhobenen Daten bietet der Tablemounted-Eye-Tracker den Vorteil, dass der Blickverlauf automatisch mit betrachteten Stimuli synchronisiert wird. Somit können die Eye-Tracking-Daten in Blickverlaufsgrafiken und -Statistiken übersetzt werden. [11, S.181] Aufgrund des in der Hochschule Reutlingen vorhandenen und zur Studie verwendeten Tablemounted-Eye-Trackers Tobii Pro X2-30 wird im Laufe der wissenschaftlichen Vertiefung ausschließlich auf seine Charakteristika eingegangen.

3.2 Der Eye-Tracking-Prozess

Der Eye-Tracking-Prozess beginnt mit einer individuellen Positionsbestimmung und Kalibrierung der Augen der Versuchsteilnehmer. Zur exakten apparativen Bestimmung des Blickzieles und zur mathematischen Rekonstruktion des Blickverhaltens sind die Positionierung des Augapfels und der Pupille sowie die Stellung des Kopfes relevant. [11, S.182] Die individuelle Kalibrierung ist notwendig, um die versuchsteilnehmerabhängigen Augen-Parameter zu optimieren und „[…] damit ein hohes Maß an Validität und Reliabilität der Messleistung sicherzustellen.“ [11, S.182]

Nach der Kalibrierung beginnt der Eye-Tracking-Prozess unter Anwendung der Cornea-Reflex-Methode. Der Eye-Tracker nimmt ein Video der Augen auf, welches die Pupillen und deren Reflexpunkt des Infrarot-Lichtstrahls auf der Hornhaut beinhaltet. Dieser Reflexpunkt wird als Cornealer Reflex bezeichnet. Dieser repräsentiert die Blickrichtung der Probanden. [7, S. 151]

Die Rohdaten nach einer Eye-Tracking-Studie liefern Informationen bezüglich der Koordinaten, der Reihenfolge der Fixationen der Augen sowie der Dauer der einzelnen Fixationen. Die Rohdaten können entweder statistisch analysiert oder als eine grafische Darstellung, die vom Eye-Tracker erstellt wird, aufbereitet werden. [10, S.12]

3.3 Gaze Plot

Gaze Plot liefert die Verweildauer und den Verlauf eines Blickes. „Die Verweildauer pro Blickobjekt (gaze duration) ist die Gesamtsumme der Zeiten, die ein Objekt oder eine Gruppe von Objekten fixiert wird.“ [7, S. 158] Jede Fixation wird als Kreis angezeigt, wobei die Größe des Kreises auf die Länge der Verweildauer hinweist [10, S. 12]. Je größer ein Kreis, desto länger ist die Fixation des Blickes auf diesem Wahrnehmungsobjekt. Die Reihenfolge der Fixationen ist nummeriert und bildet einen Verlauf der Blickwanderung (Sakkaden).

4 E-Learning-Materialien

In der Eye-Tracking-Studie werden zwei E-Learning-Materialien in Bezug auf die visuelle Aufmerksamkeit und Informationsaufnahme untersucht. Als ein weit verbreitetes Material zur statischen Informationsdarstellung werden PowerPoint-Folien gewählt. PowerPoint gilt als ein aufmerksamkeitserzeugendes Format, welches unter Verwendung von grafischen Elementen, Bildern und Stichpunktlisten Inhalte vermittelt. [25, S. 136] Dabei wird von der Einbettung von Animationen abgesehen.

Zum anderen wird für eine dynamische Darstellung der Inhalte ein Animationsfilm gewählt. Dieser präsentiert identische Inhalte anhand Bilder, die inhaltsgerecht in das Gesamtbild eingeblendet werden. Die E-Learning-Materialien präsentieren ein Thema aus der Informatik 2-Vorlesung. Es wird der Unterschied zwischen Komposition und Aggregation erörtert. Bei der Darstellung wird bewusst ausschließlich die visuelle Ebene einbezogen.

4.1 Statische Darstellung

Die statische Darstellung wird anhand der PowerPoint-Folien vorgestellt. Sie besteht aus insgesamt zwei Folien, die dem Rezipienten das Thema präsentieren. Die Folien haben eine eingestellte Überblendung nach Verlauf von 30 Sekunden. Die erste Folie ist in fünf Bereiche eingeteilt: eine übergreifende Überschrift und zwei weitere Blöcke, wobei jeder Block aus einem Textabschnitt und einer Abbildung besteht. Nach den Gesetzten der Nähe und Geschlossenheit trennt die Einteilung die zwei Begriffe innerhalb der Blöcke voneinander. [18, S. 188f] Die zweite Folie ist in drei Bereiche aufgeteilt: eine Abbildung und zwei Textabschnitte, die nach oben genannten Gesetzen aufgebaut sind.

4.2 Dynamische Darstellung

Die Umsetzung des Animationsfilmes erfolgt in Adobe Flash. Die Länge des Animationsfilmes beträgt 41 Sekunden. Die Inhalte werden als Bild- und Grafikelemente sowie als Textblöcke präsentiert. Bilder sind wirksame Mittel der Informationsdarstellung und unterstützen Verstehen und Behalten von Information. [8]

Abbildung 2: Ausschnitt aus dem Animationsfilm, Abschnitt drei

Der erste Abschnitt des Films erklärt die ‚Aggregation’ anhand von einer Grafik und Textblöcken. Nachdem die Inhalte des ersten Abschnitts ausgeblendet sind, wird im zweiten Abschnitt nach dem gleichen Schema die „Komposition“ erläutert. Der dritte Abschnitt präsentiert ein Beispiel. Dieser ist aus mehreren Bildern und erklärenden Textblöcken aufgebaut (Abb. 3).

5 Aufbau der Studie

Um die Forschungsfragen beantworten zu können, findet in der Studie eine Datenerhebung statt. Die Studie wird in der Hochschule Reutlingen durchgeführt. Als Probanden agieren Studenten der Informatik 2-Vorlesung. Die Studenten entsprechen der Zielgruppe insofern, dass sie fundierte Kenntnisse im Bereich Informatik besitzen und in der Lage sind, sich mit Thema auseinander zu setzen. Die Anzahl der Probanden in der Studie beträgt 20 Probanden, zehn für jedes E-Learning-Material. Es sind Studenten beider Geschlechter im Alter von 18 bis 30 Jahren.

Die Probanden werden in zwei Gruppen aufgeteilt. Die erste Gruppe bekommt PowerPoint-Folien, die zweite den Animationsfilm präsentiert. Die Aufgabe der Probanden ist es, das E-Learning-Material anzuschauen oder durchzulesen, um ein Verständnis über die Inhalte zu bekommen. Der Eye-Tracker verfolgt dabei den Verlauf und die Verweildauer des Blickes. Um die kontrollierte Aufnahme sicherzustellen sowie eine Entkoppelung von Aufmerksamkeit und Blickbewegungen zu vermeiden, werden die Probanden auf die abschließende Abfrage der Inhalte hingewiesen. Letztendlich ist der Zusammenhang der Informationsaufnahme mit den Ergebnissen der abschließenden Befragung zu erforschen. Hierfür werden thematische Fragen gestellt.

5.1 Technische Daten

Bei der Durchführung der Studie wird ein Tablemounted-Eye-Tracker, der Tobii Pro X2-30 benutzt. Dieser läuft mit 30HZ, das heißt, dass alle 33,3ms ein Datenpunkt erfasst wird. Die Messung im 33,3ms Intervall garantiert genaue Daten zu Blickpositionen unter realen Testbedingungen – etwa wenn sich die Nutzer bewegen. [14] Zur Analyse der Daten wird Tobii Studio verwendet. Die Software interpretiert die Daten als Fixationen und Sakkaden, wobei die letzteren zwischen den Fixationspunkten konstruiert werden.

6 Tracking-Ergebnisse

Die Ergebnisse der Eye-Tracking-Studie werden anhand von grafischen Darstellungen vorgestellt und mithilfe von Microsoft Excel analysiert. Pro E-Learning-Material werden die Daten darüber erhoben, ob ein einzelner Proband alle Elemente erfasst hat. Darüber hinaus wird festgestellt, ob ein einzelnes Element von jedem Proband erfasst wurde. Weiterhin wird angegeben, wie viele Punkte die Probanden bei der Inhaltsabfrage erreicht haben.

6.1 Statische Darstellung

Bei zehn Probanden wurden unterschiedliche Muster der Betrachtung beobachtet. Die Muster sind in den Abbildungen 3 und 4 ersichtlich.

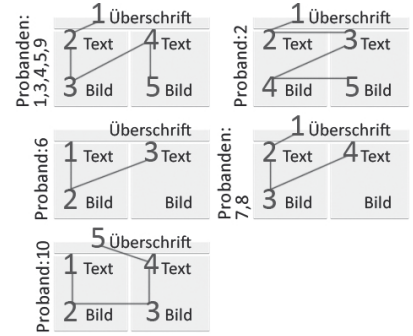

Abbildung 3:Betrachtungsmuster, Folie 1

Die Überschrift dient als zentraler Informationsbereich über das Thema und wurde von 90% der Probanden erfasst. Die Textabschnitte liefern relevante Informationen, die durch die Abbildungen verdeutlicht werden. Da die Probanden auf eine abschließende Abfrage der Inhalte hingewiesen wurden, liegt die Vermutung nahe, dass sie bewusst die textuellen Informationsbereiche zuerst abdeckten. Rayner und Pollatsek haben in ihren Untersuchungen nachgewiesen, dass sich der Aufmerksamkeitsmechanismus Wort für Wort bewegt. [5, S. 39] Das bedeutet: sobald die Probanden damit anfangen,

einen Textabschnitt zu lesen, dann muss dieser auch fertig gelesen werden. Die mäßige Aufmerksamkeit auf den Abbildungen lässt sich damit begründen, dass diese in Schwarzweiß dargestellt sind und somit wenig Salienz auslösen. [18, S. 206] Die Salienz – oder auch die Auffälligkeit eines Reizes bestimmt, ob ihm Aufmerksamkeit zugeteilt wird. [20, S. 55] 80% der Probanden gingen bei der Betrachtung nach dem von Nielsen vorgestellten F-Muster vor. [17] Das F-Pattern basiert auf der Erkenntnis, dass Rezipienten die Inhalte in ihrer gewohnten Leserichtung erkunden. Das Muster hat den Nachteil, dass die Elemente, die rechts unten stehen, nicht immer wahrgenommen werden. Dieser Sachverhalt ist in der Abbildung bei den Probanden sechs, sieben und acht ersichtlich. Dies könnte der Grund dafür sein, weshalb die Elemente neun, zehn sowie zwölf und dreizehn nur von wenigen Probanden fixiert wurden.

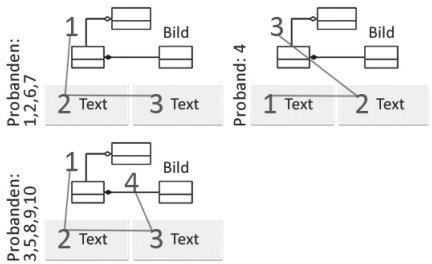

Abbildung 4:Betrachtungsmuster, Folie 2

Das Betrachtungsmuster der zweiten Folie veranschaulicht, dass die Probanden (bis auf eine Ausnahme) erst die obere Abbildung betrachteten. Die Abbildung zieht als ein großes, in der Mitte platziertes und somit hervorstechendes Element Aufmerksamkeit auf sich. [18, S. 206] Fünf von neun Probanden kehrten zu der Abbildung zurück, nachdem die Textabschnitte gelesen wurden. Wird diese Tatsache als ein Rücksprung betrachtet, so lässt es sich als eine Verarbeitungsschwierigkeit interpretieren. [21, S. 202] Eine Deutungsalternative ist das Cross-Checking-Behaviour. Raschke et al. bezeichnet dieses Verhalten als eine wiederholte Überprüfung der Richtigkeit einer Lösung, nachdem diese bereits gefunden wurde. [26, S. 15]

Insgesamt wurden in 60 Sekunden 84% der Elemente innerhalb der zwei Folien von Probanden erfasst. Die durchschnittliche Abweichung der Erfassung der Elemente liegt bei 25%. Vier von zehn Probanden haben alle Fragen richtig beantwortet. Die Rate der richtigen Antworten auf die Fragen beträgt 74%. Die Ergebnisse von Proband sieben mit 71% der fixierten Elemente und 48% richtiger Antworten sowie von Proband vier mit 100% der fixierten Elemente und 100% richtiger Antworten könnten zur folgenden Annahme führen: Die Anzahl der richtigen Antworten auf die gestellten Fragen hängt direkt von der Anzahl der wahrgenommenen und fixierten und somit vom Gehirn verarbeiteten Elemente ab. Ein direkter Zusammenhang zwischen der Fixierung der relevanten Elemente und der Beantwortung der Fragen ist aber nur bedingt vorhanden. So hat Proband sechs mit 63% die niedrigste Rate an fixierten Elementen. Bei der Abfrage hat Proband sechs mit 76% richtiger Antworten deutlich besser abgeschlossen als Proband neun mit 48%. Die Rate an fixierten Elementen beträgt bei Proband neun 92%.

6.2 Dynamische Darstellung

Die Studie hat gezeigt, dass innerhalb der drei Filmabschnitte 80% der erscheinenden Elemente im Bild eine unverzügliche Blickfixation auf sich gelenkt haben. Dies wird in der Abbildung 5 veranschaulicht. Für ein fixiertes Element wird eine „1“ notiert. Für ein nicht fixiertes Element wird eine „0“ eingetragen. Das Betrachtungsmuster ist bei allen Probanden identisch. Eine Ausnahme bildet ein Proband, der jedes Element mit einer Verzögerung, aber dennoch nach dem gleichen Muster fixiert hat. Biologische Faktoren wie Ermüdung können die Augenbewegungen beeinflussen. [19, S. 14] Somit besteht die Annahme, dass der Proband müde und deshalb nicht konzentriert war. Die Annahme wird durch die Rate der fixierten Elemente, welche mit 69% die niedrigste ist, bestätigt. Auch die Rate der richtigen Antworten von 72% bestätigt diese Aussage. Den Ausreißer bei der Beantwortung der Fragen stellt Proband fünf mit 0% der richtigen Antworten dar. Die Rate der Fixierungen liegt dabei bei 92%.

	Filmausschnitte		
Blickfixierungen der Probanden auf erscheinenden Elementen	#1	#2	#3
	1	1	1
	1	1	1
	1	1	1
	1	1	1
	1	1	1
	0	0	0
	1	1	1
	0	0	1
	1	1	1
	1	0	1
	80%	70%	90%

Abbildung 5: Blickfixationen auf Elementen (ohne Verzögerung)

Insgesamt wurden in 41 Sekunden 87% der Elemente innerhalb der drei Filmabschnitte von Probanden erfasst. Die durchschnittliche Abweichung der Erfassung der Elemente liegt bei 21%. Die Rate der richtigen Antworten auf die Fragen beträgt 78%. Zu dem am wenigsten fixierten Element zählt die „Diva ohne Schmuck“. Angenommen wird, dass sie von Probanden als nicht relevant eingestuft wurde, da sie keine für die Abfrage notwendige Information liefert. Da dieses Element dennoch eine Zustandsänderung repräsentiert, muss es in die Auswertung aufgenommen werden. Auch bei der Auswertung der dynamischen Darstellung ist zu beobachten, dass die Textabschnitte bis zum letzten Wort erfasst wurden. Die Verwendung kurzer und einfacher Sätze ermöglicht einen schnellen Sprung der Fixationen, so Rayner. [7, S. 157]

7 Diskussion

Die Ergebnisse der Studie bestätigen, dass die exogene Kontrolle anhand der dynamischen Darstellung tatsächlich Aufmerksamkeit der Probanden steuert. Die Einblendung neuer Elemente im Bild sorgt dafür, dass der Blick der Probanden aufgrund seines Orientierungsreflexes genau darauf gelenkt wird. Das ist durch das identische Betrachtungsmuster aller Probanden nachgewiesen. Auch die merkmalbasierte Selektion (Feature Selection) kann der Grund für die vermehrten Blickfixationen bei der dynamischen Darstellung sein. Diese besagt, dass sich die saliente Aufmerksamkeit auf Eigenschaften wie Farbe oder Bewegung richtet. [15], [16] Vor allem der letzte Filmabschnitt, bei dem mehrere Farben zum Einsatz kommen, hat die höchste Rate an Blickfixationen.

Die Anzahl der fixierten Elemente ist bei der dynamischen Darstellung um 3% höher als bei der statischen. Die Rate der korrekten Antworten bei der dynamischen Darstellung ist um 4% höher als bei der statischen. Die Informationsaufnahme bei der dynamischen Darstellung hat 19 Sekunden (1/3 der Gesamtzeit der statischen Darstellung) weniger gedauert und dennoch prozentuall einen, wenn auch minimal positiveren Wert erreicht (Abb. 6 und 7). Für die vorhandene Fehlerrate bei der dynamischen Darstellung kann der Orientierungsreflex verantwortlich sein. Da dieser Reflex auf die Veränderungen im Bild reagiert, kann er hohe Kosten in Form von verminderter Aufmerksamkeit verursachen. [13, S. 38] Die verminderte Aufmerksamkeit kann zu verschlechterten Gedächtnisleistungen und somit schlechteren Ergebnissen bei der Abfrage führen.

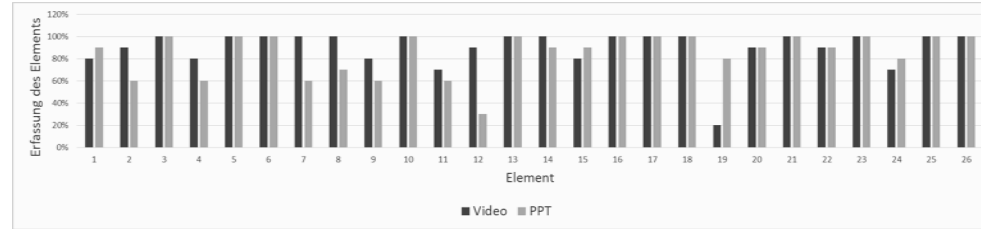

Abbildung 6: Mittelwert der Fixationen einzelner Elemente, Vergleich PPT und Animationsfilm

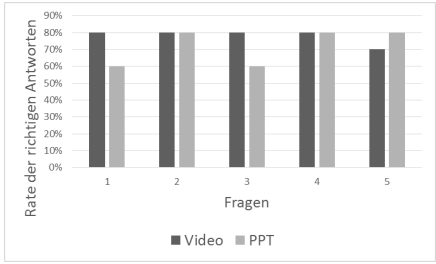

Abbildung 7: Mittelwert, Rate der richtigen Antworten, Vergleich PPT und Animationsfilm

Die Ergebnisse der Studie lassen sich mit dem Feature-Gate-Modell von Cave (1999) beschreiben. Das Modell „[…] beschreibt Informationsselektion mit Hilfe eines neuronalen Netzwerks, innerhalb dessen der Informationsfluss durch [aufmerksamkeitsabhängige] Schranken kontrolliert wird.“ [20, S. 29f] Die Schranken werden von zwei miteinander interagierenden Subsystemen gesteuert: Top-Down-System und Bottom-Up-System. Das Top-Down-System „[…] favorisiert jene Orte des visuellen Feldes, an denen sich aufgabenrelevante Merkmale befinden. […] Das Bottom-Up-System favorisiert solche Orte, an denen sich ein Merkmal stark von den umgebenen Stimuli unterscheidet.“ Bucher und Schumacher (2006) unterstützen das Modell, indem sie nahe legen, dass „[…] die Wahrnehmungsverläufe der Rezipienten als Kombination von stimulus- und goal-driven Wahrnehmung zu interpretieren sind […].“ [21, S. 158] Auch Holsanova und Holmqvist (2006) plädieren für den Einfluss der visuellen Aufbereitung der Inhalte auf die Aufmerksamkeitssteuerung. [21, S. 158]

8 Fazit

Es lässt sich zusammenfassen, dass Informationsvermittlung in Form einer dynamischen Darstellung der Inhalte sich durchaus einsetzen lässt. Sie zieht mehr Aufmerksamkeit der Rezipienten auf sich als die statische Darstellung. Obwohl dies teilweise in Folge des Orientierungsreflexes stattfindet, lassen die Ergebnisse der Befragung darauf schließen, dass die Informationsaufnahme erfolgreich war. Somit wurde in einem kürzeren Zeitraum die Aufnahme der gleichen Informationsmenge ermöglicht.

9 Literaturverzeichnis

[1] N. Hofer und W. Mayerhofer: Die Blickregistrierung in der Werbewirkungsforschung. Der Markt – Journal für Marketing. 2010

[2] S. Vögele: Werbemittel-Tests mit der Augenkamera. Königstein. 2009.

[3] H.-O. Karnath und P. Thier: Neuropsychologie. Berlin, Heidelberg, New York. 2003.

[4] C. D. Wickens und J. G. Holland: Engineering Psychology and Human Performance. Upper Saddle River, NJ: Prentice-Hall. 2000.

[5] R. Radach: Blickbewegungen beim Lesen. Psychologische Aspekte der Determination von Fixationspositionen. Münster, New York. 1996.

[6] R. J. Gerrig und P. G. Zimbardo: Psychologie. Pearson Deutschland GmbH. 18. Aufl. 2008.

[7] G. Rickheit, T. Herrmann, W. Deustch: Psycholinguistics: Ein internationales Handbuch. Walter de Gruyter. 2003

[8] Shu, N.C. (1988): Visual Programming. Van Nostrand Reinhold Company, New York, NY.

[9] A. Butz, A. Krüger: Mensch-Maschine-Interaktion. Walter de Gruyter GmbH. 2014.

[10] J. Nielsen und K. Pernice: Eyetracking web usability. Addison Wesley Longman. 2009.

[11] G. Nufer und V. Ambacher: Eye Tracking als Instrument der Werbeerfolgskontrolle. Hrgb.: C. Rennhak und G. Nufer. 2012.

[12] A. Duchowski: Eye Tracking Methodology: Theory and Practice. Springer Verlag. 2007.

[13] W. Leven: Blickverhalten von Konsumenten: Grundlagen, Messung und Anwendung in der Werbeforschung. Springer Verlag. 2013

[14] Eye-Tracker Tobii Pro X2-30. http://www.tobiipro.com/product-listing/tobii-pro-x2-30. Accessed 26 October 2015

[15] A. M. Tresiman und G. Gelade: A feature-integration theory of attention. Cognitive Psychology. 1980. Seite 97-136.

[16] M. Corbetta, F. M. Miezin, S. Dobmeyer, G. L. Shulman, S. E. Petersen: Attentional modulation of neural processing of shape, color, and velocity in humans. Science 248: Seite 1556 – 1559. 1990

[17] J. Nielsen: F-Shaped Pattern For Reading Web Content. Jakob Nielsen’s Alertbox. 2006

[18] S. Thesmann: Einführung in das Design multimedialer Webanwendungen. Vieweg+Teubner. 2010.

[19] F. Thömke: Augenbewegungsstörungen: Ein klinischer Leitfaden für Neurologen. Georg Thieme Verlag. 2008

[20] I. Rosendahl: Der Einfluss auffälliger Reize auf die Aufmerksamkeit. Herbert Utz Verlag. 2001.

[21] S. Geise: Extended Paper Eyetracking in der Kommunikations- und Medienwissenschaft. Studies in Communication and Media. Seite 149-263. 2011

[22] G.W. McConkie, M.D. Reddix und D. Zola: Chronometric analysis of language processing during eye fixation. Annual Meeting of the Psychonomic Society. Boston, MA. 1985

[23] G. D. Rey: Methoden der Entwicklungspsychologie: Datenerhebung und Datenauswertung. BoD. 2012

[24] M. A. Just und P. A. Carpenter: A theory of reading: From eye fixations to comprehension. Psychological Review. Seite 329-354. 1980

[25] J. Koch und J. Sydow: Organisation von Temporalität und Temporärem. Springer-Verlag. 2013

[26] M. Raschke, T. Blascheck, M. Richter, T. Agapkin, T. Ertl: Visual Analysis of Perceptual and Cognitive Processes. IVAAP. 2014

Analyse von One-Time Password Loginmethoden zur Dezimierung von Fremdzugriffen

Steffen Schellig
Reutlingen University
Steffen.Schellig@Student.
Reutlingen-University.DE

Abstract
In dieser Arbeit werden One-Time Passwörter analysiert und mit anderen Authentisierungsmethoden verglichen. Weiter werden Risiken und Schwachstellen der einzelnen Methoden besprochen und wie sich die Nutzer davor schützen können. Weiter werden in diesem Zusammenhang die verschiedenen Authentisierungsmethoden nach unterschiedlichen Kriterien verglichen. Danach wird durch eine Nutzerstudie geprüft, was dem durchschnittlichen Nutzer wichtiger ist, wenn es um die Sicherheit seiner Onlineaccounts geht, Zeit oder Sicherheit. Zuletzt soll diese Arbeit mit einer Zusammenfassung der Ergebnisse und einem Ausblick auf zukünftige Schritte im Bereich der Authentisierung abgeschlossen werden.

Schlüsselwörter
Authentication, 2-Factor, One-Time Password, biometric, comparison.

Betreuer Hochschule: Prof. Dr. Michael Tangemann
Hochschule Reutlingen
Michael.Tangemann@Reutlingen-University.de

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

CR-Kategorien
D.4.6 Security and Protection – Authentication

1 Einführung
In der heutigen Zeit geschieht immer mehr online, egal ob Soziale Medien, E-Mails, Onlinebanking, Geschäftliches oder bei der Informationssuche. Dabei ist man auch nichtmehr Ortsgebunden am Rechner zuhause, sondern kann von überall mit Hilfe des Smartphones Onlineinhalte abrufen. Die meisten vertraulichen Daten sind dabei durch Passwörter geschützt oder sollten dies sein. Nun hatte, wie in Abb.1 zu sehen ist, im Jahr 2014 der durchschnittliche Internetnutzer 5,5 Social-Media-Accounts. Hierzu kommen meist noch geschäftliche Account sowie solche für Online-Banking.

Bezieht man nun noch einen Online-Banking-Account mit ein, besaß ein 16-24 Jähriger im Jahr 2014 ca. sieben Online-Accounts und somit potenzielle Angriffsziele. Außerdem sollten bei diesen sieben Accounts die Passwörter nicht zu häufig wiederholt werden, weshalb man von sechs Passwörtern mit jeweils Groß-/Kleinbuchstaben, Sonderzeichen und Zahlen ausgehen kann. Diese hohe Anzahl an

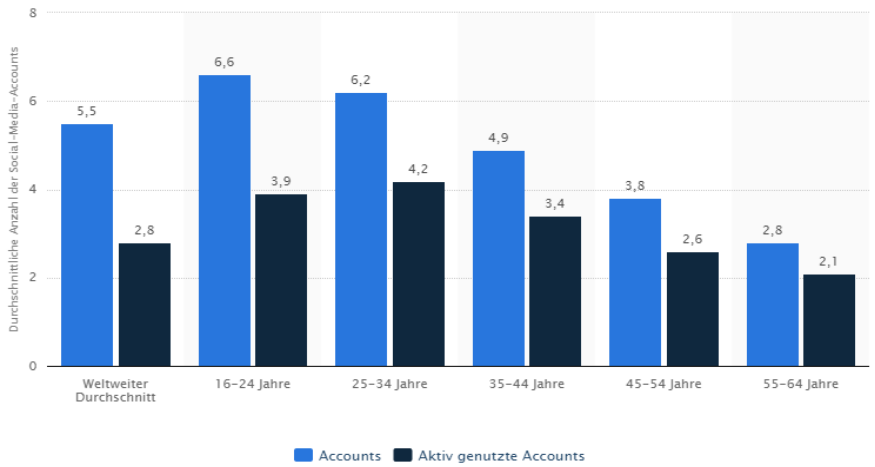

Abbildung 1: Social-Media-Accounts 2014 [18]

verschiedenen Passwörtern bringt einen zusätzlichen Verwaltungsaufwand und ein zusätzliches Risiko mit, da man sie meist mit dem passenden Account niederschreibt, um den Überblick zu bewahren.

An diesem Punkt setzt folgende Arbeit an, die einen Überblick über verschiedene One-Time Passwörter gibt und durch verschiedene Vergleiche und Analysen deren Einsetzbarkeit prüft. Unter einem One-Time Passwort soll in dieser Arbeit ein Passwort verstanden werden, welches auf eine Anfrage hin generiert wird und nur für kurze Zeit und einmalige Nutzung gültig ist.

1.2 Ziele

Ziel dieser Arbeit soll ein Vergleich und eine Analyse verschiedener One-Time Passwort Varianten sein. Hierbei soll die Funktionsweise, die Benutzerfreundlichkeit und die Sicherheit überprüft werden, um zu evaluieren, wie gut diese Loginmethoden gegenüber den herkömmlichen abschneiden. Weiter soll in einer Nutzerumfrage geklärt werden, ob eine erhöhte Sicherheit bei Logins vorgezogen wird, selbst wenn dadurch ein erhöhter Bedienungsaufwand entsteht.

1.1 Vorgehensweise und Gliederung

Zu Beginn dieser Arbeit wird ein Einblick in den aktuellen Stand der Wissenschaft gegeben, welcher sich mit den verschiedenen One-Time Passwort Methoden und der dazugehörigen Theorie befasst. Am Beispiel des TAN-Verfahrens sollen die beschriebenen Methoden dann nochmals erklärt und gefestigt werden. Danach folgen die Risiken dieser Loginmethoden, welche sich in Schwachstellen und Angriffsmöglichkeiten unterteilen. In Kapitel vier folgt eine Nutzerstudie zur Nutzerfreundlichkeit verschiedener Anwendungen und inwieweit ein möglicher Mehraufwand angenommen wird. Am Ende der Arbeit folgen ein Fazit über die vorherigen Ergebnisse und ein Ausblick auf zukünftige Arbeiten, welche in diesem Bereich verfolgt werden könnten.

2 Stand der Wissenschaft

In diesem Kapitel sollen verschiedene Theorien und Methoden beschrieben werden, welche für One-Time Passwörter benutzt werden. Danach soll am Beispiel des TAN-Verfahrens, wie es Banken verwenden, der Ablauf eines solchen Mechanismus genauer aufgezeigt werden. Zuletzt sollen diese Methoden mit den herkömmlichen Passwörtern verglichen werden, um eine Abgrenzung herzustellen.

2.1 One-Time Passwortarten

Die Methoden nach denen One-Time Passwörter arbeiten, lassen sich in vier grundlegende Kategorien unterteilen. Dabei handelt es sich um mathematische Algorithmen, Smart Cards, zeitlich synchronisierte Tokens und SMS.

2.1.1 Mathematische Algorithmen

Im Jahr 1981 schlug Lamport [7] die erste One-Time Passwort Authentisierung vor, und nutzte dafür eine „one-way hash chain“. [8] Lamport hatte zwei Ideen, welche er für diese Art der Authentisierung umsetzen wollte. Zum einen sollte sein System nicht die Passwörter direkt speichern, sondern Werte, die aus einer bestimmten Funktion und dem Passwort erstellt wurden. Dies sollte es verhindern, dass Passwörter aus dem System ausgelesen werden können, da die Umrechnung von einem gespeicherten Wert zurück in das Passwort nur schwer möglich wäre. Zum anderen sollte nicht nur ein einzelnes Passwort, sondern eine Sequenz von beispielsweise 1000 Passwörtern genutzt werden. Hierbei müsste jedes genutzte Passwort gestrichen werden, um ein Abfangen und Wiederbenutzen von Passwörtern auszuschließen. [7] Diese Art der Authentisierung hätte jedoch einen relativ hohen Wartungsaufwand, da das Ausgangspasswort nach allen 1000 Authentisierungen neu gewählt werden müsste.

2.1.2 Smart Card

Eine zweite Variante der One-Time Passwörter setzt auf die Nutzung von sogenannten Smart Cards. Als einer der ersten reichte der Franzose Roland Moreno im Mai 1976 ein Patent ein, welches im Verlauf zur Smart Card führte. Er beschreibt sie in seinem Patent als:

„A portable independent electronic object designed for storing and transferring data confidentially intended for being coupled to a data transfer device; [...] intended for comparing the enabling data contained in the store with a confidential code supplied by the rightful owner of the portable object and introduced into the portable object by way of the said transfer device.“

[19]

Somit stellt die Smart Card ein unabhängiges und portables Speicher- und Transportmedium dar, welches sensible Daten enthalten kann und durch Authentisierung an rechtmäßige Besitzer ausgibt. Weiter können diese Karten nur mit jeweils passenden Lesegeräten gelesen oder beschrieben werden, was auch einen der größten Nachteile dieser Authentisierungsform ausmacht. Da das Mitführen dieser Gegenstände für die Nutzer eine größere Last wäre, beschränkt sich der Gebrauch der Smart Cards hauptsächlich auf den Heimbereich.

2.1.3 Zeitlich synchronisierte Tokens

Unter zeitlich synchronisierten Tokens versteht man eine Authentisierung mithilfe eines Zweitgerätes und der Passwortberechnung aus der aktuellen Uhrzeit. Diese Methode wurde von M’Raihi et al. [24] im RFC 6328 festgehalten und beschreibt die Nutzung eines Hardwaretokens zur Generierung von zeitlich synchronisierten Passwörtern. Die Tokens können hierbei auch moderne Smartphones mit passender Applikation sein, wie es beispielsweise bei „Google Authenticator“ der Fall ist.

Bei dieser Art der Authentisierung wird die Uhrzeit meist in eine gewisse Anzahl an Zeitslots unterteilt. Jedem dieser Slots sind dann die gültige Uhrzeit und ein Passwort zugeordnet, welches nur in diesem bestimmten Zeitraum besteht. Durch eine Anfrage des Tokens an den Server wird dann die Uhrzeit beider Geräte verglichen und ein Passwort an den Token gesandt, mit welchem

sich der Benutzer dann beim Serverdienst anmelden kann. [1]

2.1.4 Short Message Service

Als vierte Methode für One-Time Passwörter führen Liao et al. [8] Authentisierungsmethoden via Short Message Service (SMS) auf. Diese Methode kann heute jedoch auch auf passende Smartphoneapplikationen ausgeweitet werden, da die SMS immer mehr durch Apps wie WhatsApp, Threema etc. ersetzt wird. Die Funktionsweise und die meisten bestehenden Risiken bleiben jedoch gleich.

SMS und andere Kommunikationsdienste verfahren nach dem „Best effort" Prinzip, was in diesem Fall bedeutet, dass der Anbieter des Dienstes weder die Ankunft der Nachricht, noch die Übertragungszeit dieser garantiert. Dies spielt im Zusammenhang mit One-Time Passwörtern eine ernstzunehmende Rolle, da ein generiertes Passwort nur eine begrenzte Zeit lang gültig sein sollte. [8]

2.2 TAN-Verfahren

Die Abkürzung TAN steht hierbei für Transaktionsnummer und beschreibt einen Code, der benötigt wird, um eine Transaktion erfolgreich zu tätigen. H. Kubicek und G. Diederich stellen in ihrem Buch „Sicherheit im Onlinebanking" [16] vier verschiedene TAN-Verfahren vor, welche von den Banken zu Transaktionszwecken genutzt werden. Auf diese vier Arten soll hier kurz eingegangen werden, um ein Grundverständnis für diese unterschiedlichen TAN-Verfahren zu fördern.

2.2.1 TAN-Liste

Bei der TAN-Liste handelt es sich um eine, per Brief zugestellte Liste mit einer gewissen Anzahl an TANs. Nach dem Login im Benutzerkonto der Bank wird für den Abschluss einer Transaktion eine dieser Nummern eingegeben. Jede TAN kann hierbei nur einmalig verwendet werden und verliert nach Verwendung an Gültigkeit. Sobald sämtliche TANs auf der Liste verbraucht sind oder auf Anforderung des Nutzers wird eine Liste mit neuen TANs per Post zugesandt. [16]

2.2.2 iTAN

Die iTAN kann als sichere TAN-Liste gesehen werden und steht für indizierte TAN. Hierbei bekommt jede TAN auf der Liste eine Nummer zugeteilt. Bei der Onlinetransaktion wird nun von der Bank eine dieser Nummern vorgegeben, womit nur die dazugehörige TAN für diese Transaktion gültig ist. Wie bei der TAN-Liste sind alle Tans nur einmal gültig und eine neue Liste wird auf Anforderung zugesandt. [16]

Diese Variante kann grob mit den zeitlich synchronisierten Tokens aus Kapitel 2.1.3 verglichen werden, da zu einer Transaktion nur eine spezielle TAN genutzt werden kann.

2.2.3 SMS-TAN

Die SMS-TAN ist auch unter Kürzeln wie smsTAN, mTAN oder mobileTAN bekannt und arbeitet ohne eine Liste, die dem Nutzer vorab zugesandt wurde. Bei der SMS-TAN muss der Nutzer sein Mobiltelefon bei der Bank registrieren, welches dann für Transaktionen benötigt wird. Nach Absenden der Transaktionsdaten im Onlinebereich der Bank erhält der Nutzer eine zeitlich begrenzte TAN, welche er zum Abschluss der Transaktion im Onlinebereich eingeben muss. [16]

2.2.4 ChipTAN

Beim chipTAN-Verfahren wird ein zusätzliches Gerät von der Bank benötigt. Durch Einführen der Bankkarte in dieses Gerät wird eine TAN errechnet und auf dem Display dieses TAN-Generators angezeigt. Diese TAN wird dann vom Nutzer im Onlineformular eingetragen und für die Transaktion abgeschickt. [16]

Abbildung 2: Verschiedene Online-banking-Verfahren [17]

In Abb. 2 ist ein Überblick über die Nutzung verschiedener Onlinebanking-Verfahren von 2011 aufgezeigt. Diese Statistik kann sich aber in den vergangenen Jahren deutlich verändert haben, da in diesem Zeitraum auch die Anzahl der Smartphones drastisch zunahm, womit die mTAN mehr ins Blickfeld der Nutzer gerückt sein könnte. Dieser Punkt soll in Kapitel 5 mituntersucht werden, da es an passenden, aktuellen Statistiken mangelt. Bezüglich der Sicherheit der einzelnen Verfahren kann gesagt werden, dass mit Aufkommen der ChipTAN und der mTAN viele Fortschritte gemacht wurden. Jedoch sind auch diese beiden Methoden nicht unüberwindbar.

2.3 Vergleich zu verschiedenen Passwortmethoden

Was genau unterscheidet nun die One-Time Passwörter von verschiedenen anderen Passwortarten und bringen die One-Time Passwörter wirklich eine Verbesserung der Sicherheit?

Beim Vergleich der angesprochenen Authentisierungsmethode mit herkömmlichen Passwörtern zeigt sich, dass One-Time Passwörter einen kleineren Verwaltungsaufwand besitzen, jedoch etwas mehr Zeit und Arbeitsschritte zum Login benötigen. Im Gegensatz zu den herkömmlichen Authentisierungsverfahren muss sich kein Passwort mehr gemerkt werden, woraus sich ein geringerer Verwaltungsaufwand ergibt. Die Arbeitsschritte und der zeitliche Aspekt stamen vom Ablauf der Authentisierung bei One-Time Passwortmethoden. Hierbei muss zuerst das Passwort angefordert werden, um sich danach mit diesem einzuloggen. Das Abrufen des Passworts kann hierbei auch auf einem externen Gerät geschehen, was diese Methode wiederum etwas aufwändiger für den Nutzer macht, da er dieses Gerät stets mitführen muss, um sich zu authentisieren.

Im Vergleich zu verschiedenen biometrischen Authentisierungsmethoden wie beispielsweise Fingerscans oder Gesichtserkennungen benötigt man beim One-Time Passwort nicht in allen Fällen eine spezielle Hardware. In manchen Fällen wie beim externen TAN-Gerät kann diese Unterscheidung jedoch nicht getroffen werden. Vergleicht man die

Fehleranfälligkeit dieser beiden Authentisierungsmethoden kann gesagt werden, dass die biometrischen Methoden schon durch leichte Veränderungen des zu scannenden Objekts fehlschlagen können. Hierzu könnten beim Fingerscann schon kleine Wunden zu einer Nicht-Erkennung des Eigentümers führen.

Eine weitere Authentisierungsmethode die zum Vergleich herangezogen werden kann, ist die 2-Faktor Authentisierung. Diese ähnelt der One-Time Passwortmethode, wie sie beispielsweise in Kapitel 2.1.4 beschrieben ist. Hierbei wird ein zusätzliches Gerät zum erfolgreichen Login benötigt. Dies kann ein Smartphone, ein Tablet oder ähnliches sein. Bei dieser Variante wird ein Verifizierungscode an dieses externe Gerät versandt, welcher dann zum erfolgreichen Login am Computer benötigt wird. Dieser zusätzliche Code an sich kann wie ein One-Time Passwort verstanden werden, welcher kurzzeitig und für nur einen Login gültig ist. Im Gegensatz zur One-Time Passwortmethode ist dieser Code jedoch ein Zusatz und man benötigt weiterhin einen Benutzernamen samt normalem Passwort.

3 Risiken

In diesem Kapitel sollen verschiedene Risiken, Schwachstellen und Angriffsmöglichkeiten gegen verschiedene Passwortmethoden gezeigt werden, wobei ein besonderes Augenmerk auf die One-Time Passwörter gelegt wird.

3.1 Vergleichskriterien

Die unterschiedlichen Vergleichskriterien, die nachfolgend aufgeführt werden, sollen zur Analyse der verschiedenen Risiken und Schwachstellen der einzelnen Authentisierungsmethoden herangezogen warden, um einen Vergleich zu schaffen und eine grobe Ordnung der einzelnen Passwortarten hinsichtlich dieser Kriterien zu geben.

3.1.1 Äußere Faktoren

Unter äußeren Faktoren sollen im Folgenden Einflüsse verstanden werden, welche nicht in direktem Zusammenhang mit dem Login stehen, sondern beispielsweise als Kommunikationsweg dienen. Unter diese Faktoren fallen zum einen externe Kommunikationswege wie in Kapitel 2.1.4 für die SMS-Methode. Zum anderen sollen hierunter extern benötigte Geräte Fallen, wie zum Beispiel ein TAN-Generator. Bei diesen Vergleichskriterien soll untersucht werden, inwieweit bei der Übertragung der Daten Fehler auftreten können, oder inwiefern der Nutzer durch zusätzliche Geräte eingeschränkt wird.

3.1.2 Arbeitsschritte

Als Arbeitsschritte werden im Folgenden die Schritte beschrieben, die ab dem Seitenaufruf, bis hin zum erfolgreichen Login vom Nutzer zu erledigen sind. Dabei werden sämtliche Geräte die für diesen Vorgang benutzt werden zusammengefasst um einen Überblick zu schaffen, welchen Aufwand der Nutzer bei den beschriebenen Authentisierungsmethoden hat. Die Arbeitsschritte sollen unter Kapitel 5 auch in einer Nutzeranalyse untersucht werden, um herauszufinden inwieweit Nutzer bereit sind, einen höheren Aufwand einzugehen, um sich sicherer mit Onlinekonten zu verbinden.

3.1.3 Vorgangsdauer

Das Kriterium der Vorgangsdauer beschreibt nachfolgend die ungefähre zeitliche Dauer vom Beginn des Loginvorgangs bis zu dessen erfolgreichem Anschluss. Dabei werden alle Kommunikationswege mitberücksichtigt, die bei der Anmeldung zu bestreiten sind. Es wird lediglich ein Richtwert als beschrieben, da die Sendezeiten bei SMS oder E-Mails für 2-Faktor Authentisierungen schwanken können und keine feste zeitliche Größe in Anspruch nehmen.

3.1.4 Subjektive Sicherheit

Die subjektive Sicherheit soll ein Richtwert sein, welcher den Aufwand beschreibt, den ein Angreifer für eine Übernahme des Accounts eingehen müsste. Dabei wird auf verschiedene Angriffsszenarien Rücksicht genommen. Dieser Richtwert ist rein subjektiv, da er in keiner Einheit gemessen werden könnte und sehr Situationsabhängig ist. Dieses Kriterium wird dennoch berücksichtig, um Lesern bei der Auswahl, für Sie geeigneter Authentisierungsmethoden zu unterstützen.

3.2 Schwachstellen & Angriffsmöglichkeiten

Im folgenden Kapitel werden die Schwachstellen der unterschiedlichen Authentisierungsmethoden analysiert und beschrieben, sowie passende Angriffsmöglichkeiten aufgezeigt, welche diese Schwachstellen ausnutzen. Hierbei soll zuerst auf die üblichen Passwörter eingegangen werden, bevor speziellere Authentisierungsmethoden wie biometrische Authentisierung, 2-Faktor Authentisierung und die besprochenen One-Time Passwörter in Augenschein genommen werden.

3.2.1 Passwörter

Die üblichen Passwörter können als wohl einfachste Form der Authentisierung beschrieben werden und bieten eine breite Angriffsfläche. Um diese Passwörter zu umgehen, benötigt man nicht unbedingt hochmoderne Hard- und Software oder ein tiefergehendes Verständnis der Informatik. Viele Angriffe lassen sich selbst von Laien durchführen und benötigen keine spezielle Hard- oder Software.

In diesem Abschnitt sollen nachfolgend verschiedene Angriffsmöglichkeiten wie beispielsweise Brute-Force und Social Engineering besprochen werden, um aufzuzeigen, wie einfach es sein kann, diese Passwörter zu umgehen oder an passende Anmeldeinformationen zu gelangen. Des Weiteren soll ein kurzer Hinweis gegeben werden, wie man sich gegen diese Angriffsarten besser absichern kann.

3.2.1.1 Brute-Force

Beim Brute-Force („Rohe Gewalt“) oder auch „exhaustive search“ („auslaugende Suche“) Angriff werden sämtliche Buchstaben und Zahlenfolgen nacheinander auf Erfolg getestet, was unter Umständen eine lange Zeit in Anspruch nehmen kann. [15] Diese Angriffsart kann schon von Privatpersonen eingesetzt werden, welche über nur geringe informationstechnische Grundlagen verfügen, da viele selbsterklärende Programme im Internet leicht zu bekommen sind. Durch größtenteils eingeschränkte Hardwaremöglichkeiten dieser Privatpersonen können hier gute Passwörter jedoch schon einen guten Schutz bieten.

3.2.1.2 Social Engineering

Unter Social Engineering („soziale Manipulation“) versteht man das Erschließen persönlicher Informationen durch soziale Interaktion mit dem Opfer. Dadurch erhoffen sich Angreifer an Informationen zu gelangen, welche es ihnen ermöglichen, in die Accounts der Opfer einzudringen. Da viele Benutzer Passwörter oder Sicherheitsfragen aus ihrem näheren Umfeld nutzen, um ihre Accounts zu schützen, können Angreifer durch soziale Plattformen wie Facebook leicht an diese Daten gelangen. Auch die direkte soziale Interaktion in Unterhaltungen kann ausgenutzt warden, um an Accountinformationen zu gelangen. Hierbei setzen die Angreifer auf verschiedene psychologische Verhaltensmodelle, welche Bezuidenhout et al. in ihrem Paper "Social engineering attack detection model: SEADM." [2] gut beschreiben.

Einen Schutz gegen die genannten Angriffsmöglichkeiten auf Passwörter bieten bereits viele Internetseiten durch das Einfügen sogenannter Captchas beim Login, welche menschliche Benutzer von automatischen Loginvorgängen trennen sollen, indem abgebildete Wörter eingegeben werden. Ein weiterer Schutzmechanismus, welcher von den Nutzern selbst getroffen werden kann, ist der Aufbau und die Länge des Passworts. Das Bundesamt für Sicherheit in der Informationstechnik empfiehlt hierfür ein Passwort, welches aus min zwölf Zeichen besteht und Groß-, Kleinbuchstaben,

Sonderzeichen und Zahlen enthält. Außerdem sollte es nicht in Wörterbüchern vorkommen oder bekannten Mustern entsprechen, wie beispielsweise ‘12345678’. Des Weiteren sollten keine persönlichen Daten wie Namen von Freunden, Familienmitgliedern oder Haustieren oder Geburtsdaten genommen werden, um Social Engineering vorzubeugen. [21]

Ein gutes Beispiel des BSI ist ein selbsterdachter Satz, welcher als Eselsbrücke für ein Passwort dienen soll. Hierbei werden dann nur die Anfangsbuchstaben des Satzes genommen und Buchstaben wie „i“ und „l“ durch „1“ ersetzt. So entsteht aus dem Satz: ‘Morgens stehe ich auf und putze mir meine Zähne drei Minuten lang.’ Folgendes Passwort: ‘Ms1a&pmmZ3M1’. [21]

3.2.2 Biometrische Authentisierung

Die biometrischen Authentisierungsverfahren wie Finger- oder Irisscans sind etwas komplizierter zu umgehen als die üblichen Passwörter und erfordern meist spezielle Hardware oder tieferes Verständnis über ihre Funktionsweise und informationstechnischen Methoden. Jedoch gibt es auch für diese Authentisierungsmethode verschiedene Wege, um sich illegal Zugriff zu verschaffen. Nachfolgend sollen einige davon aufgezählt und beschrieben werden.

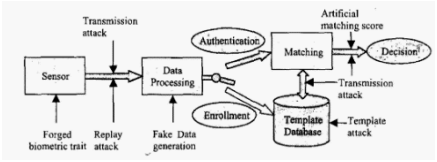

Abbildung 3: Angriffe auf biometrische Authenstisierungsverfahren [13]

In Abb. 3 sind verschiedene Angriffsmöglichkeiten mit samt ihren Einstiegspunkten in biometrischen Systemen zu sehen.

3.2.2.1 Spoof attack

Bei der sogenannten Spoof Attack („Täuschungsangriff“) gibt ein Angreifer vor, der legitime Benutzer zu sein, der sich auf einem System einloggen möchte. Dieser Angriff auf biometrische Authentisierungsmethoden verwendet gefälschte biometrische Kopien des eigentlichen Benutzers, um den Sensor aus Abb. 3 zu täuschen. Zu diesen gefälschten Kopien zählen beispielsweise hochauflösende Aufnahmen der Iris, gefälschte Fingerabdrücke, Masken oder Sprachaufnahmen. [13]

Daraus kann man schließen, dass ein Angreifer, der sich dieser Methode bedient eine spezielle Hardware besitzt und einen direkten Kontakt zum Opfer benötigt, um an die erforderlichen Originaldaten zu gelangen.

3.2.2.2 Replay attack & Transmission attack

Replay attacks und Transmission attacks können im groben zusammengefasst werden, da sich ihre Funktionsweisen ähneln. Der Einstiegspunkt dieses Angriffs ist hierbei nicht der Sensor sondern der Kommunikationsweg danach.

Bei der „Replay attack“ werden Daten abgefangen, welche vom Sensor zum Prozessor geschickt werden. Der Angreifer nutzt diese Daten dann, um sich selbst Zugriff zu verschaffen, indem er den Sensor umgeht und die abgefangenen Daten direkt an den Prozessor schickt.

Die „Transmission attack“ kann an verschiedenen Kommunikationswegen einsteigen, wie man in Abb. 3 sehen kann. Hierzu sendet der Angreifer manipulierte Daten weiter, erstellt infizierte, künstliche Sensorergebnisse oder injiziert gefälschte Antworten auf die Sensorabfrage. [13]

3.2.2.3 Template attack

Bei der „Template attack“ versuchen die Angreifer die biometrischen Templates, die auf dem System abgelegt sind, zu verändern, zu stehlen oder eigene Templates einzufügen und greifen somit die in Abb. 3 gezeigte „Template Database“ an. Sollte es ihnen gelingen Templates vom System zu stehlen, könnten diese dazu ausgenutzt warden, um das biometrische System zu entschlüsseln und so Fälschungen zu erstellen, die vom System erkannt werden. Des Weiteren wäre durch den Diebstahl solcher Templates die biometrische Einzigartigkeit der Nutzer nicht mehr gewährleistet, da ihre biometrischen Daten Fremden zur Verfügung stehen, welche Sie beliebig oft fälschen könnten. [13]

Der Schutz gegen Angriffe und Diebstähle kann bei dieser Authentisierungsmethode nicht wirklich vom Nutzer beeinflusst werden, sondern liegt in den Händen der Anbieter. Beispielsweise können durch verschlüsselte Übertragungen innerhalb des Systems die „Transmission attack“ erschwert werden. Auch würde eine verschlüsselte Speicherung der Templates die „Template attack“ um einiges komplizierter werden lassen, da die Angreifer die Templates nun erst noch entschlüsseln müssten, um an die biometrischen Daten zu gelangen. Laut Xiao [13] wäre der beste Schutz eine manipulationssicheres Gerät für die biometrische Authentisierung zu verwenden, da dies sämtliche Risiken deutlich einschränken würde. Die einzige Möglichkeit das System dann noch zu täuschen, wären die „Spoof attacks“, die auch vom Benutzer nur schwer abzuwehren sind.

3.2.3 2-Faktor Authentisierung

„Two-factor authentication isn’t our savior. It won’t defend against phishing. It’s not going to prevent identity theft. It’s not going to secure online accounts from fraudulent transactions. It solves the security problems we had 10 years ago, not the security problems we have today.“ [10]

Mit diesen Worten eröffnet Bruce Schneier, der zu diesem Zeitpunkt CTO der Counterpane Internet Security Inc. war, im April 2005 seine ACM-Veröffentlichung über die 2-Faktor Authentisierung.

Zwei Möglichkeiten, die 2-Faktor Authentisierung zu umgehen, wären beispielsweise ein Man-in-the-Middle Angriff oder der Angriff mithilfe eines Trojaners.

3.2.3.1 Man-in-the-Middle Angriff

Bei dieser Angriffsmethode schleicht sich der Angreifer zwischen das Opfer und die Internetseite, die der Geschädigte besuchen wollte. Beispielsweise setzt der Angreifer eine gefälschte Homepage der Bank auf und lenkt den Nutzer auf diese. Dort fängt er die Benutzeranmeldung des Nutzers ab und loggt sich mit diesen direkt auf der Original Bankseite ein. Hierbei würde das sich ändernde Passwort der 2-Faktor Authentisierung einfach vom Angreifer übernommen und verwendet werden. Wenn der Angreifer die gefälschte Webseite hierbei dem Original gut anpasst, fällt dem Nutzer möglicherweise erst sehr spät oder gar nicht auf, dass er nicht auf der originalen Seite ist. [10]

3.2.3.2 Angriff mit Trojaner

Bei einem Angriff auf diese Authentisierungsmethode mit Hilfe eines Trojaners ist der Angreifer gezwungen, diesen Trojaner zuerst auf dem PC oder SMartphone des Opfers zu installieren. Dies könnte beispielsweise durch das Besuchen infizierter Websites oder das Öffnen infizierter E-Mails geschehen. Ist der Trojaner einmal installiert, könnte der Angreifer beispielsweise eine Session des Nutzers übernehmen, sobald dieser sich in seinen Bankaccount einloggt. Der Nutzer selbst würde hiervon nicht mitbekommen und würde es frühestens beim nächsten Kontrollieren seiner Kontoauszüge bemerken. [10]

Aber auch gegen diese Arten der Angriffe kann man sich als Nutzer wappnen. Als wohl wichtigster Punkt kann hierbei gesagt werden, dass man niemals suspekte Links oder Internetseiten öffnen sollte. Ein weiterer Punkt ist das Prüfen der Verbindung zu Homepages, die man öfters besucht, wie Bankaccounts oder soziale Medien. Hier

sollte man darauf achten, dass die Seite stets über HTTPS erreicht wird und gültige Zertifikate besitzt, was meist durch die verschiedenen Internetbrowser als Grün hinterlegte URL-Leiste signalisiert wird und somit schnell zu identifizieren ist. Um den Befall durch Trojaner einzuschränken, sollte auch hier darauf geachtet werden, weder dubiose Internetseiten noch E-Mails, beziehungsweise deren Anhänge zu öffnen. Vor allem bei angeblichen Rechnungen, Mahnungen oder Vorladungen welche über E-Mail als .ZIP Datei versandt werden, sollte man vorsichtig sein. Sollte es jedoch ein Trojaner geschafft, haben sich zu installieren, helfen schon meist die Standard Virenscanner oder sogenannte „Rescue-Disks“, welche aus einer sicheren Umgebung heraus starten und das System auf Schadsoftware untersuchen.

3.2.4 One-Time Passwörter

Durch die ähnliche Funktionsweise wie bei der 2-Faktor Authentisierung, unterscheiden sich die Angriffsmöglichkeiten auf One-Time Passwörter nicht wirklich von denen aus Kapitel 3.1.3. Hauptsächlich wird auch hier das Session Hijacking genutzt, wie es im vorherigen Kapitel beschrieben wurde. [23]

Eine weitere Möglichkeit für Angreifer wäre die Übernahme des zum Erhalt des One-Time Passworts genutzten Accounts. Falls der Nutzer sich die Passwörter einer Internetseite beispielsweise auf seine E-Mailadresse schicken lassen würde und sich der Angreifer Zugriff auf diese E-Mailadresse verschaffen könnte, wäre auch die Authentisierung auf der Internetseite in Gefahr.

3.3 Aktuelle Beispiele

In diesem Kapitel sollen zur Veranschaulichung der zuvor beschriebenen Risiken zwei Beispiele beschrieben werden, welche in den letzten Jahren aufkamen und von denen viele Privatnutzer betroffen waren.

3.3.1 Geodo

Bei Geodo handelt es sich um einen Online-Banking-Trojaner, welcher im Mai 2014 bekannt wurde und sich über E-Mail verbreitete. Hierbei wurden E-Mails von angeblichen Telekommunikationsdienstleistern mit Anhängen oder Links verschickt, die zum Download des Trojaners führten. Sobald der Trojaner auf einem PC installiert war, manipulierte dieser Transaktionen im Online-Banking und spähte die Zugangsdaten zu E-Mail-Konten aus, welche dazu genutzt wurden, um den Trojaner weiter zu verbreiten. [20]

3.3.2 iBanking

Das zweite Beispiel, welches seit Mitte 2013 bekannt ist, bedient sich der Smartphones, um Online-Banking zu manipulieren. Hierbei sind vorwiegend Android-Smartphones betroffen und Online Transaktionen, welche mit Hilfe des mTAN-Verfahrens arbeiten. Zu Beginn wird der PC eines Opfers mit einem Trojaner infiziert, welcher anspringt, sobald man sich bei seinem Online-Banking-Account anmeldet. Nach der Anmeldung wird die Webseite der Bank so manipuliert, dass Handydaten angefordert warden, um fortzufahren. Das Opfer erhält hierauf eine SMS mit dem Link zu einer App, welche mit dem iBanking-Trojaner infiziert ist. Sind PC und Smartphone infiziert, startet die Schadsoftware auf dem PC eine Überweißung. Der iBanking-Trojaner auf dem Smartphone fängt die mTAN, welche zum erfolgreichen Abschließen der Transaktion benötigt wird, ab und sendet diese an den PC. Die dort installierte Schadsoftware übermittelt nun die mTAN an die Bank und schließ somit die eigene Geldüberweißung erfolgreich ab. [20]

3.4 Vergleich nach genannten Kriterien

Vergleicht man nun die vier genannten Authentisierungsmethoden anhand der beschrieben Kriterien, Risiken und Schwachstellen, können folgende Schlüsse gezogen werden:

Passwörter an sich bieten einen sehr schnellen und direkt Weg zum Login, sind jedoch die Variante, bei der die Accountdaten am schnellsten in falsche Hände geraten können. Hierzu tragen verschiedene Möglichkeiten für Angreifer bei, wie beispielsweise Brute-Force Angriffe, die Veröffentlichung von sogenannten Rainbow Tables oder Social Engineering. Bei einem guten Passwort können jedoch viele dieser Risiken eingedämmt werden.

Biometrische Authentisierungsmethoden bieten hingegen einen besseren Schutz, sind jedoch unter den genannten Arten die einzigen, die bisher im Onlinebereich noch nicht wirklich zum Einsatz kommen. Zwar haben Angreifer auch hier mehrere Möglichkeiten, den Login zu umgehen, jedoch erfordern diese ein höheres Verständnis der Informatik und deutlich bessere Hardware. Der Aufwand und die Dauer dieser Methode sind ähnlich einzustufen wie bei den Passwörtern. Jedoch herrscht hier eine höhere Fehleranfälligkeit durch Veränderungen der biometrischen Daten.

Die 2-Faktor Authentisierung kann insoweit als sicherer eingestuft werden, da mehrere Geräte bereits infiziert sein müssten, um an die Accountinformationen zu gelangen. Beispielsweise müsste sich Schadsoftware auf dem Smartphone und dem Computer befinden, um einige Angriffe erfolgreich abzuschließen. Des Weiteren sind bei einer höheren kritischen Betrachtung des Internets durch die Nutzer solche Schadprogramme gut zu vermeiden.

Bei den One-Time Passwörtern kann die Sicherheit subjektiv auch höher eingeschätzt werden als bei vielen Passwörtern, da durch das Generieren neuer Passwörter bei jeder Anmeldung viele Sicherheitsrisiken entfallen. Jedoch kann hier der Aufwand und die Dauer des Logins höher eingestuft werden. Auch können bei Anforderung des Passworts über SMS oder Internet zusätzliche Kosten für den Nutzer entstehen. Der Aufwand des Niederschreibens oder Einprägens von Passwörtern entfällt jedoch.

4 Nutzerstudie

In diesem Kapitel wird die durchgeführte Nutzerbefragung hinsichtlich Account-sicherheit und den verschiedenen genutzten Passwortarten beschrieben und analysiert.

4.1 Hypothesen

Die erste Hypothese leitet sich aus einem Artikel des DIVSI zum Thema „PRISM und die Folgen: Sicherheitsgefühl im Internet verschlechtert“ her, indem es heißt, dass seit den Snowden-Enthüllungen und dem damit zusammenhängenden Abhörskandal das Sicherheitsgefühl der Deutschen im Internet gesunken ist. [22] Dies führt zu folgender Hypothese: „Wenn ein Nutzer die Sicherheit seines Accounts gering einschätzt, dann ist er eher gewillt einen zusätzlichen Loginaufwand einzugehen, um die Sicherheit

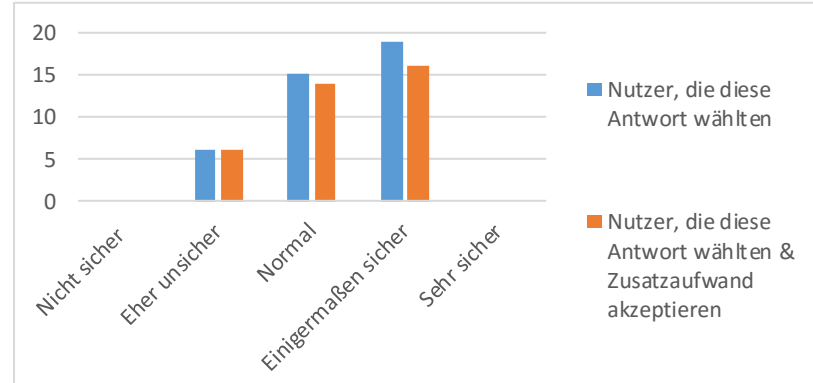

Abbildung 4: Eingeschätzte Sicherheit - Akzeptanz zusätzlichen Aufwandes

zu verbessern.“ Diese Hypothese führt zum einen zur Frage, wie man die Sicherheit seiner Accounts einschätzt. Zum anderen zu Fragen bezüglich des Einsatzes eines zusätzlichen Gerätes oder eines erhöhten Aufwandes um die Sicherheit der Accounts zu steigern. Die Frage nach einem zentralen Login steht diesen beiden Fragen entgegen, da sich hiermit nicht nur der Aufwand senken würde, sondern auch die Möglichkeit mehrere Accounts zur selben Zeit zu verlieren, erheblich steigt.

Zusätzliche Fragen zur Nutzung verschiedener Authentisierungsmethoden oder der subjektiven Einschätzung zu den größten Sicherheitsrisiken der Accounts sollen einen Überblick über die Akzeptanz unterschiedlicher Loginmöglichkeiten geben und aufzeigen, was den Teilnehmern am meisten Sorgen bereitet, wenn es um die Sicherheit ihrer Daten geht.

4.2 Durchführung

Die Befragung fand innerhalb von zwei Tagen an verschiedenen Universitäten in Baden-Württemberg statt. Hierbei wurden ausschließlich Studenten angesprochen und befragt, weshalb die Studie auch als selektive Stichprobe gesehen werden sollte. Die Teilnehmer waren aus verschiedenen Studienrichtungen im Alter zwischen 17 und 30 Jahren und hatten unterschiedliche Kenntnisse im Bereich der Informatik. An der Befragung nahmen 40 Studenten teil, wovon 17 männlich und 23 weiblich waren.

4.3 Ergebnisse

Betrachtet man die erste Hypothese und die zugehörigen erhobenen Daten zeigt sich zum einen, dass der Großteil der Befragten einen zusätzlichen Aufwand oder ein Zweitgerät nur für spezielle Accounts nutzen würden. Hierzu gaben 35% an, für geschäftliche Konten ein Zweitgerät, zusätzlichen Aufwand oder beides in Kauf zu nehmen. Für Bankkonten waren es hingegen schon 75% der Personen, die den oben genannten Mehraufwand eingehen würden. Unterscheidet man nun zwischen einem Zweitgerät oder einem höheren Aufwand durch Anforderung eines Einmalpasswortes an beispielsweise eine E-Mail-Adresse, liegt die Akzeptanz des Zweitgerätes leicht unter der anderen Variante. So würden 75% der Befragten ein Zweitgerät zur Sicherheit hinzuziehen, einen Mehraufwand durch Passwortanfrage hingegen 90%. Diese Werte sind dadurch begründet, dass Mehrfachantworten möglich waren und einzelne Personen auch mit beidem einverstanden waren.

Das Diagramm in Abb. 4 zeigt das Verhältnis der Sicherheitseinschätzung zur Akzeptanz eines Mehraufwandes. Dabei symbolisieren die Blauen Balken die Anzahl aller Nutzer, welche die unten beschrieben Accountsicherheit gewählt hatten. In Orange wird die Anzahl der Nutzer abgebildet,

welche passend zur Sicherheitseinschätzung einen Zusatzaufwand oder ein Zweitgerät

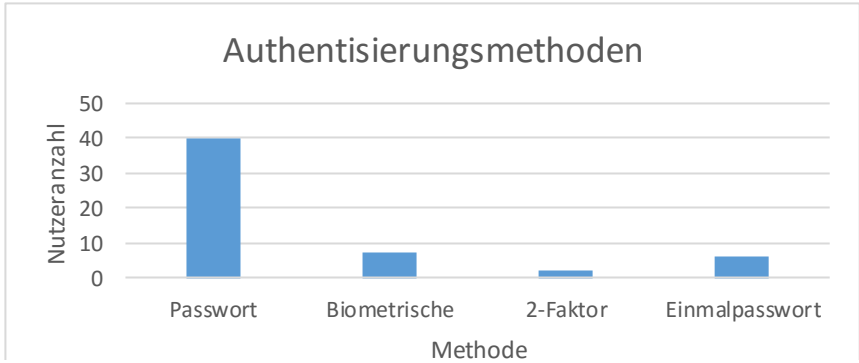

Abbildung 5: Genutzte Authentisierungsmethoden

zum Login umsetzen würden. Obwohl die Testgruppe recht klein war, lässt sich erkennen, dass die Akzeptanz für zusätzlichen Loginaufwand abnimmt, je sicherer die Personen ihre Accounts einschätzen. Im Gegensatz zum beschriebenen Artikel aus Kapitel 4.1 [22], welcher die Sicherheitseinschätzung von Accounts in Deutschland relativ niedrig einstufte, kann hier gesagt werden, dass der Großteil der Befragten ihre Accounts eher sicher einschätzen.

Wenn man die Akzeptanz der verschiedenen Authentisierungsmethoden betrachtet, kann festgestellt werden, dass alle befragten Nutzer Passwörter nutzen und sich nur manche zusätzlich mit anderen Methoden absichern, wie in Abb. 5 zu sehen ist. So nutzen beispielsweise ca. 17,5% der Befragten zusätzliche noch biometrisch Authentisierungsverfahren, 15,0% sichern sich mit Einmalpasswörtern weiter ab und nur ca. 5,0% nutzen die 2-Faktor Authentisierung. Diese Ergebnisse haben nicht zwangsweiße etwas mit der Akzeptanz der Authentisierungsmethoden zu tun, sondern könnten auch auf höheren Unbekanntheitsgrad zurückzuführen sein, was sich durch diverse Zwischenfragen der Befragten zeigte.

Bei der Frage nach den größten Sicherheitsrisiken gestanden sich ca. 40,0% der Teilnehmer selbst die Schuld zu, indem sie angaben, entweder zu einfache oder oft dieselben Passwörter zu benutzen. Die zweitgrößte Gefahrenquelle sahen ca. 25,0% der Befragten in Hackern oder Angriffe auf Datenbanken mit gespeicherten Accountinformationen. Von den 40 Befragten Personen nannten fünf die NSA oder Amerika als Gefahr die Sicherheit ihrer Daten, was mit großer Wahrscheinlichkeit auf die Snowden-Enthüllungen zurückzuführen ist.

Bei den beschriebenen Ergebnissen sollte jedoch festgehalten werden, dass diese noch nicht auf Signifikanz geprüft wurden sie in weiteren Auswertungen deshalb auf eine solche überprüft werden sollten.

5 Fazit und Ausblick

In diesem Kapitel wird die Arbeit kurz zusammengefasst, um wichtige Ergebnisse nochmals darzulegen. Auch wird ein kleiner Leitfaden für Nutzer zum sichereren Umgang mit Onlineaccounts beschrieben. Zu Letzt wird ein Ausblick auf zukünftige Arbeiten gegeben und eine Prognose, wie sich die Passwortsituation entwickeln könnte.

5.1 Zusammenfassung und Bewertung der Ergebnisse

In den vorherigen Kapiteln wurden verschiedene Passwortarten besprochen und auf ihre Sicherheit untersucht. Weiter wurden diese verglichen und es fand eine subjektive Einschätzung statt, in welchem Verhältnis Aufwand, Sicherheit, Dauer und andere Faktoren dieser Authentisierungsmethoden zusammenhängen. Die durchgeführte Nutzerstudie zur Accountsicherheit und der Akzeptanz genannter Authentisierungsmethoden zeigte, dass ein Großteil der Authentisierungsmethoden noch wenig genutzt werden, obwohl viele Nutzer einen Zusatzaufwand eingehen würden, um ihre Accounts sicherer zu machen. Dies könnte an fehlendem Wissen über die Verfügbarkeit solcher Methoden liegen. Im Folgenden sollen noch einmal verschiedene Möglichkeiten aufgeführt werden, die Sicherheit seiner Accounts zu steigern.

5.2 Leitfaden zur Accountsicherheit

In diesem Kapitel werden einige Ratschläge und Möglichkeiten genannt, wie man sich als Nutzer gegen viele Angriffe schützen kann.

Für die Auswahl der richtigen Passwörter sollte man darauf achten, dass man möglichst nicht dieselben Passwörter auf verschiedenen Accounts verwendet. Auch vom Verändern seiner üblichen Passwörter durch Hinzufügen einer „1“ oder eines „!“ am ende eines Wortes, sollte abgesehen werden. Des Weiteren sind Wörter, die im Duden

vorkommen, Familien- und Städtenamen oder einfache Alphanumerische Folgen meist recht einfach zu umgehen. Ein gutes Passwort, welches man sich einfach merken kann sind die, in Kapitel 3.2.1 beschriebenen Passwortsätze. [21]

Falls man über eine hohe Anzahl an Accounts und Passwörtern verfügt, kann man sich diese auch ohne Bedenken notieren. Jedoch sollte hierbei darauf geachtet werden, wo man sich seine Accountdaten niederschreibt. Von einfachen Textdateien auf dem Desktop des Computers oder von Papieren direkt auf dem Schreibtisch ist eher abzuraten. Verschafft sich ein Angreifer Zugang zum Computer, so sind die Accountdaten nicht lange sicher, wenn sie sich nur in einer unverschlüsselten Datei befinden. Falls man sich seine Daten dennoch notieren möchte, wäre ein verstecktes Notizbuch die bessere Wahl. Auch für Computer und Smartphone gibt es bessere Alternativen als die Textdatei, wie beispielsweiße verschiedene Programme, welche Passwörter verschlüsselt abspeichern und nur durch ein Masterpasswort zu öffnen sind.

Eine weitere Möglichkeit, um Accounts besser zu schützen, sind die One-Time Passwörter, welche in dieser Arbeit behandelt wurden und heutzutage von vielen Anbietern zur Authentisierung angeboten werden. Hierbei wird das Passwort zufällig generiert und auf eine hinterlegte E-Mail oder ein Smartphone geschickt. Diese Möglichkeit zur Authentisierung bietet einen guten Schutz, da Angreifer in diesem Fall Schadsoftware auf den Computer und das Smartphone bringen müssten, um an Accountdaten zu gelangen.

Im Allgemeinen sollte man neben guten Passwörtern auch darauf achten, seine Geräte von Schadsoftware freizuhalten. Hierbei helfen oft schon ein aktueller Virenschutz und eine gewisse Vorsicht gegenüber Links und Dateien in E-Mails. Es sollte beispielsweise davon abgesehen werden, Anhänge des Dateityps „.exe“ oder „.zip“ zu öffnen, falls man den Absender nicht sicher kennt.

5.3 Offene Fragen und künftige Arbeit

One-Time Passwörter sind weiterhin am Kommen und werden inzwischen von einigen großen Anbietern zur Authentisierung bei Onlineaccounts genutzt. Ob sich dieser Trend durchsetzt oder ob neue Anmeldeverfahren kommen werden, bleibt abzuwarten. Auch die Verbesserung bereits bestehender Authentisierungsmethoden könnte einen Fortschritt bringen, indem neue Sicherheitsmöglichkeiten oder Verschlüsselunsalgorithmen implementiert warden.

6 Literaturverzeichnis

[1] Aljareh, Salem and Kavoukis, Anastasios. „Efficient Time Synchronized One-Time Password Scheme to Provide Secure Wake-Up Authentication on Wireless Sensor Networks“. International Journal Of Advanced Smart Sensor Network Systems (IJASSN), Vol 3, No.1, January 2013

[2] Bezuidenhout, Monique, Francois Mouton, and Hein S. Venter. "Social engineering attack detection model: SEADM." Information Security for South Africa (ISSA), 2010. IEEE, 2010.C

[3] Cheng, Fred. "Security attack safe mobile and cloud-based one-time password tokens using rubbing encryption algorithm." Mobile Networks and Applications16.3 (2011): 304-336.

[4] Goyal, Vipul, et al. "The N/R one time password system." Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on. Vol. 1. IEEE, 2005.

[5] Kim, Hyun-Chul, et al. "A design of one-time password mechanism using public key infrastructure." Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on. Vol. 1. IEEE, 2008.

[6] Korte, Ulrike et al. „Datenschutzfreundliche Authentisierung mit Fingerabdrücken“. BSI – Datenschutz und Datensicherheit https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Studien/BioKeys/DatenschutzfreundlicheAuthentisierungmitFingerabdruecken.pdf?__blob=publicationFile

[7] Lamport, Leslie. „Password Authentication with Insecure Communication“. Communications oft he ACM, Vol 24, Nr. 11, (1981): 70-72

[8] Liao, Kuan-Chieh, et al. "A one-time password scheme with QR-code based on mobile phone." INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on. IEEE, 2009.

[9] Rayes, Mohamed Omar. "One-time password." Encyclopedia of Cryptography and Security (2011): 885-887.

[10] Schneier, Bruce. "Two-factor authentication: too little, too late." Commun. ACM 48.4 (2005): 136.

[11] Vaidya, Binod, et al. "Robust one-time password authentication scheme using smart card for home network environment." Computer Communications 34.3 (2011): 326-336.

[12] Wang, Ding, and Ping Wang. "Understanding security failures of two-factor authentication schemes for real-time applications in hierarchical wireless sensor networks." Ad Hoc Networks 20 (2014): 1-15.

[13] Xiao, Qinghan. "Security issues in biometric authentication." Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC. IEEE, 2005.

[14] Häder, Michael. „Empirische Sozialforschung – Eine Einführung“. Springer Wiesbaden (2015) ISBN: 978-3-531-19674-9. DOI 10.1007/978-3-531-19675-6

[15] Knudsen, Lars R., and Matthew JB Robshaw. "Brute force attacks." The Block Cipher Companion. Springer Berlin Heidelberg (2011). ISBN: 978-3-642-17341-7

[16] Kubicek, Herbert, and Günther Diederich. "Sicherheit im Online-Banking.". Springer Wiesbaden (2015) ISBN: 978-3-658-09959-6. DOI: 10.1007/978-3-658-09960-2

[17] Statista - Welche Verfahren setzen Sie zur Abwicklung von Online-Banking-Transaktionen ein? http://de.statista.com/statistik/daten/studie/219644/umfrage/verfahren-zur-abwicklung-von-online-banking-transaktionen/ (zuletzt geprüft: 06.10.2015)

[18] Statista - Anzahl der Social-Media-Accounts pro Internetnutzer weltweit nach Alter im 4. Quartal 2014 http://de.statista.com/statistik/daten/studie/383689/umfrage/anzahl-der-social-media-accounts-pro-internetnutzer/ (zuletzt geprüft: 03.09.2015)

[19] USPTP Patent Full-Text And Image Database. United States Patent 4,092,524, Moreno, May 30, 1978. http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=4092524.PN.&OS=PN/4092524&RS=PN/4092524 (zuletzt geprüft: 11.09.2015)

[20] BSI – Bundesamt für Sicherheit in der Informationstechnik. „Die Lage der IT-Sicherheit in Deutschland 2014“

(2014) https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2014.pdf;jsessionid=E0997B2F53A4ED9A125F65D94C2E9BC8.2_cid368?__blob=publicationFile

[21] BSI – Bundesamt für Sicherheit in der Informationstechnik. „Passwörter“ https://www.bsi-fuer-buerger.de/BSIFB/DE/MeinPC/Passwoerter/passwoerter_node.html (zuletzt geprüft: 18.10.2015)

[22] DIVSI – Deutsches Institut für Vertrauen und Sicherheit im Internet. „PRISM und die Folgen: Sicherheitsgefühl im Internet verschlechtert“ https://www.divsi.de/prism-und-die-folgen-sicherheitsgefuehl-im-internet-verschlechtert/

[23] Haller, Neil, et al. A one-time password system. No. RFC 2289. 1998. https://www.rfc-editor.org/rfc/pdfrfc/rfc2289.txt.pdf

[24] M'Raihi, David, et al. „Totp: Time-based one-time password algorithm“. No. RFC 6238. 2011. http://www.rfc-editor.org/rfc/rfc6238.txt

Konzeption einer Systemarchitektur zur Verbesserung der Performance bei der Produktsuche in Onlineshops *

Lukas Schmitt
Reutlingen University

lukas.schmitt@student.reutlingen-university.de

Abstract

Eine schnelle Suche in e-Commerce Systemen ist zwingend notwendig, um potenzielle Kunden nicht durch lange Ladezeiten oder fehlerhafte Ergebnisse zu verlieren. Deshalb wird in dieser Arbeit eine Systemarchitektur entworfen, welche die Reaktionszeit der Shopsuche erhöht. Dies wird am Beispiel der internetstores GmbH durchgeführt. Die bestehende Architektur wird dabei analysiert und modifiziert. Als Grundlage dafür wird eine allgemeingültige Systemarchitektur entworfen. Der Vergleich der beiden Systeme zeigt, dass allein durch die prototypische Implementierung der neuen Systemarchitektur eine Leistungssteigerung von bis zu 70% erreicht werden kann.

Schlüsselwörter

Datenbank, Produktsuche, Systemarchitektur, e-Commerce

CR-Kategorien

C.0 [**Computer Systems Organization**]: Modeling of computer architecture

---

* Betreuer Hochschule: Prof. Dr. Peter Hertkorn, Hochschule Reutlingen, Peter.Hertkorn@Reutlingen-University.de
Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

1 Einleitung

Die Firma internetstores GmbH betreibt diverse Onlineshops, wie z.B. fahrrad.de, bruegelmann.de oder campz.de. Allein in den deutschen Shops werden zu Spitzenzeiten bis zu einer Millionen Suchanfragen am Tag bearbeitet. Die Datenquelle für diese Shopsysteme besteht im Moment aus einem Zusammenschluss von Caches und Datenbanken. Durch die aktuelle Umsetzung benötigt eine Änderung an einem Produkt im schlimmsten Fall bis zu einer Stunde und 20 Minuten, bis diese im Onlineshop verfügbar und über die Suchfunktion sichtbar ist. Dies ist vor allem der kommerziellen Lösung Factfinder zuzuschreiben, da diese nur jede Stunde ihre Informationen mit der SQL-Datenbank abgleicht. Die SQL-Datenbank eines Shopsystems wird wiederum vom ERP-System befüllt. Die Produktsuche eines Shops ist an das Factfinder-System gebunden, was bedeutet, dass ein Suchergebnis ebenfalls veraltete Daten enthalten kann. Das Factfinder-System wiederum liefert lediglich Produkt-Schlüssel, anhand derer im Redis-Cache die passenden Datenobjekte abgerufen werden können. Die Arbeit soll eine Möglichkeit aufzeigen, wie der Produkt-Update-Zyklus minimiert und die Reaktionszeit der Suche erhöht werden kann.

Die restliche Arbeit ist wie folgt aufgebaut: Im nächsten Kapitel werden ähnliche Arbei-

ten zu diesem Thema vorgestellt. Anschließend wird die Umgebung der internetstores GmbH kurz analysiert und die Anforderungen für die neue Architektur werden definiert. Im nachfolgenden Kapitel wird die allgemein gültige Systemarchitektur beschrieben, welche als Grundlage verwendet wird. Im nächsten Kapitel wird beschrieben, wie die Architektur speziell für die internetstores angepasst wurde. Im vorletzte Kapitel werden die definierten Anforderungen gegenüber der neuen Architektur geprüft und die Leistungsdaten des neue und alten Systems verglichen. Im letzte Kapitel befindet sich das Fazit zur Arbeit.

2 Ähnliche Arbeiten

Im Bereich der Architektur-Optimierung und der Verbesserung der Suche gibt es bereits einige Arbeiten.
Wang et al.[20] haben in ihrer Arbeit ihren Focus auf die Verbesserung der Suche gelegt. In ihrem Ansatz verbinden sie die Such-Engine Lucene mit einer Oracle-Datenbank. Durch die Verwendung der Such-Engine wollen sie vor allem die Volltextsuche verbessern. Die Such-Engine indexiert die Inhalte der Datenbank, um spätere Suchanfragen verarbeiten zu können. Mit dieser Lösung können sowohl Bereichssuchen, als auch Suchen auf spezielle Datumsfelder durchgeführt werden.

Schlachter et al.[19] konzentrieren sich bei ihrem Ansatz konkret auf Informationsportale im Web. Sie erläutern, dass diese Portale ihre Informationen aus vielen verschiedenen Datenbanken beziehen, welche unterschiedliche Datenformen bereitstellen, z.B. Messdaten, Metadaten oder Berichte zu bestimmten Themen als Volltext. Diese Informationen werden auf der Seite über reguläre Navigationen bzw. eine Suchmaske erreichbar gemacht. Ziel von Schlachter et al. ist es, diese Informationen zu bündeln und zu logischen Objekten zusammen zu fassen. Dies soll über eine Such-Engine realisiert werden, da Informationsportale mit einer großen Menge an Daten arbeiten.

Jun Bai [4] will die Log-Daten, die in einem Unternehmen anfallen, in Echtzeit sammeln und analysieren. Gerade in großen Unternehmen fallen Millionen von Log-Informationen an, welche für Analysen verwendet werden können, um so wertvolle Informationen zu erhalten. Echtzeit-Analysen sind in einem Oracle-Cluster laut Bai nur schlecht möglich, weshalb eine andere Lösung gefunden werden muss, welche bei großen Datenmengen besser skaliert. Google und Co. setzen für ihre Datenverarbeitung längst nicht mehr auf relationale Datenbanken, sondern auf NoSQL-Systeme, wie z.B. BigTable. Die Datenbank HBase von Apache Hadoop ist die Open-Source-Variante von Bigtable. Mit diesem System können auch große Datenmengen einfach gespeichert werden. Der Nachteil dieses Systems ist jedoch, dass es nur einen Schlüssel gibt, den Spaltenschlüssel. Bai löst dieses Problem durch die Verwendung einer Such-Engine. Der Defakto-Standard von Unternehmen ist die Such-Engine Lucene. Diese ist jedoch nicht als verteiltes System konzipiert. Die Such-Engine Elasticsearch baut auf Lucene auf und ist für verteilte Systeme ohne zusätzliche Konfiguration geeignet. Somit verbindet Bai HBase und Elasticsearch, um Log-Daten zu speichern und zu analysieren.

3 Analyse der Umgebung

3.1 Das Produkt

Das Produkt-Objekt im Shopsystem repräsentiert eine Ansammlung von kaufbaren Artikeln. Diese Artikel werden Variationen genannt. Die verschiedenen Variationen lassen sich zu einer logischen Einheit bündeln - dem Produkt - da ihre Grundeigenschaften identisch sind. Die Variationen eines Produktes unterscheiden sich lediglich in ihrer Ausprägung von bestimmten Eigenschaften wie z.B. Farbe oder Größe. So hat beispielsweise ein Produkt P vier verschiedene Größen (S, M, L, XL), wobei jede Größe eine Variation des Produktes darstellt. Die Variationen können sich über alle konfigurierbaren Eigenschaften erstrecken. Eine Eigenschaft ist dann konfigurierbar, wenn sie mindestens zwei unterschiedliche Werte annehmen kann. Ein Wert für eine konfigurierbare Eigenschaft wird Ausprägung genannt. Wenn eine Eigenschaft für ein Produkt und deren Variationen immer gleich ist, ist sie nicht konfigurierbar und somit eine Grundeigenschaft. Ein Produkt kann somit n Variationen besitzen, wobei n das mathematische Produkt der Anzahl der Ausprägung der konfigurierbaren Eigenschaften ist. Beispielweise hat ein Produkt mit den Variationen Größe und Farbe mit den Ausprägungen S, M, L, XL für die Größe und rot, grün, blau für die Farbe, 12 verschiedene Variationen.[1]

Im Intenetstores-Shopsystem gibt es pro Produkt maximal eine konfigurierbare Eigenschaft. Alle weitere Eigenschaften, die theoretisch als Variation für dieses Produkt in Frage kommen, werden wiederum als eigenständiges Produkt geführt. Für das Beispiel von oben würde es somit nicht nur ein Produkt geben, sondern vier Produkte - für jede Farbe ein Produkt. Jedes Produkt hat die konfigurierbare Eigenschaft Größe mit den Ausprägungen S, M, L, XL.[1]
Für die Darstellung der Produkte gibt es zwei Bereiche im Shopsystem: Die Produktübersicht und die Produktdetailseite. Die Produktdetailseite stellt ein Produkt mit seinen verschiedenen Variationen dar. Dazu gehören auch Bewertungen durch Kunden, Produktbilder und verschiedene Zusatzinformationen, die für den Kunden von Interesse sein könnten. Auf der Produktübersichtseite werden alle Produkte zu einer bestimmten Kategorie dargestellt. Diese können über verschiedene Filteroptionen wie z.B. Größe, Farbe oder Preisrahmen eingeschränkt werden, um das gewünschte Produkt zu finden. Die Produktfilter werden anhand der Produkte in der Datenbank erstellt, was bedeutet, dass alle Produkte einer Kategorie ausgelesen werden müssen, um die passende Produktübersicht darzustellen. Der Vorteil ist, dass die Filteroptionen nicht extra

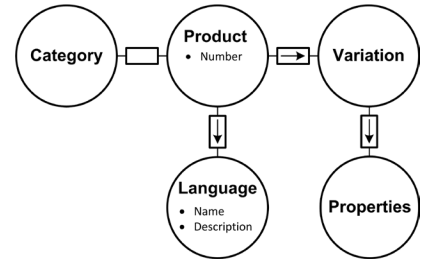

Abbildung 1: Wertestrukturdiagramm des Produkts[18]

gepflegt werden müssen, sondern diese sich direkt aus den Produkten ergeben und somit immer so aktuell wie die Produkte selbst sind. Dadurch können leere Filteroptionen vermieden werden. Das sind genau solche Optionen, zu denen keine Produkte mehr hinterlegt sind. Ein Produkt ist mindestens einer Kategorie zugeordnet, kann jedoch auch zu mehreren Kategorien gehören.[1]

Die Abbildung des Produktes in der relationalen Datenbank umfasst 22 Tabellen, die miteinander verknüpft sind. Um ein Produkt vollständig darstellen zu können, werden alle 22 Tabellen benötigt. Eine Abfrage für eine Liste von Produkten, z.B. für ein Suchergebnis, dauert entsprechend lang, da eine Abfrage mit vielen Tabellenverknüpfungen sehr rechenintensiv ist.[3][8] Die Datenbank fast ca. 500.000 Produkte, aus denen sich knapp eine Millionen Variationen ergeben - stetig wachsend. Abbildung 1 zeigt den Aufbau des Produkts als Wertestrukturdiagramm. Aus Gründen der Übersichtlichkeit wurde darauf verzichtet, alle 22 Tabellen abzubilden.[1]

3.2 Systemarchitektur

Im sogenannten Backend-System befindet sich das ERP-System. Hier werden die Produktdaten von Mitarbeitern eingepflegt und verwaltet. Im Frontend, dem eigentlichen Shopsystem, befindet sich eine relationale Datenbank (MariaDB), die ebenfalls die

Produktdaten enthält, sowie weitere Informationen, die für den Betrieb des Shops benötigt werden. Diese ist die Hauptdatendank, auf die das Shopsystem zugreift. Das Factfinder-System spiegelt die Produktdaten für die Suche. Es ist so konfiguriert, dass es nur die Produktschüssel zurückliefert. Daher ist das Ergebnis der jeweiligen Factfinder-Suche eine Menge von Produktschlüsseln. Neben der Produktschlüssel liefert das Factfinder-System auch die nötigen Filterinformationen für die Produktübersichtsseite. Da dem Factfinder-System alle Produktinformationen vorliegen, kann dieses die nötigen Filteroptionen zusammentragen und für eine bestimmte Kategorie an das Shopsystem übergeben.[1]

Da das Factfinder-System alle Treffer liefert und nicht nur die ersten x Treffer, wäre die Menge der Daten potentiell sehr groß, wenn Factfinder alle Informationen zu einem Produkt liefern würde und nicht nur den Schlüssel. Daher wird zusätzlich eine Key-Value-Datenbank eingesetzt, um die eigentlichen Produktinformationen abzurufen. Dabei handelt es sich um eine Redis-Datenbank (auch Redis-Cache). Die Produktobjekte, die im Shopsystem verwendet werden, werden serialisiert in der Redis-Datenbank gespeichert. Mit den Produktschlüsseln können genau diese serialisierten Objekte abgerufen und im System de-serialisiert werden. Auf die Produktinformationen in der relationalen Datenbank wird nur ein Mal zugegriffen, genau dann, wenn das Produkt nicht in der Redis-Datenbank vorhanden ist und das Objekt erst neu erstellt werden muss. Es wird anschließend im Redis-Cache gespeichert, sodass bei der nächsten Anfrage auf das Produkt, nicht mehr auf die relationale Datenbank zugegriffen werden muss. Die Abbildung 2 zeigt das Aufbaudiagramm der beschriebenen Architektur.[1]

Das Shopsystem wurde von der Internetstores selbst entwickelt und basiert auf dem PHP-Framework Symfony2. Der Zugriff auf die relationale Datenbank findet über den ORM-Layer Doctrine statt. Die von Doctrine erstellten Objekte, welche den Inhalt der Datenbank abbilden, sind genau die, die im Redis-Cache vorgehalten werden. Somit kann das Shopsystem mit den Doctrine-Objekten arbeiten, ohne sie tatsächlich aus der relationalen Datenbank bekommen zu haben.[1]

3.3 Produkt-Update

Eine Produktänderung im ERP-System benötigt im aktuellen System bis zu einer Stunde und 20 Minuten, bis die Änderung für den Kunden sichtbar und in der Suche berücksichtigt wird. Das ERP-System exportiert ca. alle 10 Minuten die neusten Änderungen in XML-Dateien. Diese werden wiederum alle 10 min von der sogenannten Shop-API abgerufen. Die Shop-API ist eine Sammlung von Programmelementen, welche die Informationen in den Datenbanken verwalten. Jedoch ist der Cronjob des ERP-Systems nicht synchron mit dem der Shop-API, weshalb nicht gewährleistet ist, dass diese direkt nacheinander durchgeführt werden. So kann es dazu kommen, dass die XML-Dateien erst kurz vor dem nächsten Export des ERP-Systems von der Shop-API abgerufen werden. Damit ergibt sich eine maximale Zeitspanne von bis zu 20 Minuten, bis die Daten tatsächlich von der Shop-API abgerufen werden.[1]

Die Shop-API liest die Änderung aus den XML-Dateien und wendet diese auf die relationale Datenbank an, sodass die Produkte an dieser Stelle wieder auf dem neusten Stand sind. Anschließend werden diverse Cache-Tabellen erzeugt, welche für einen schnelleren Zugriff auf Produktkategorien benötigt werden. Da das Shopsystem jedoch seine Produktinformationen nicht direkt aus der relationalen Datenbank bezieht, sondern aus dem Redis-Cache, werden weiterhin die veralteten Daten ausgeliefert. Daher werden nun die Redis-Einträge der geänderten Produkte neu erzeugt, sodass das Shopsystem die aktuellsten Informationen

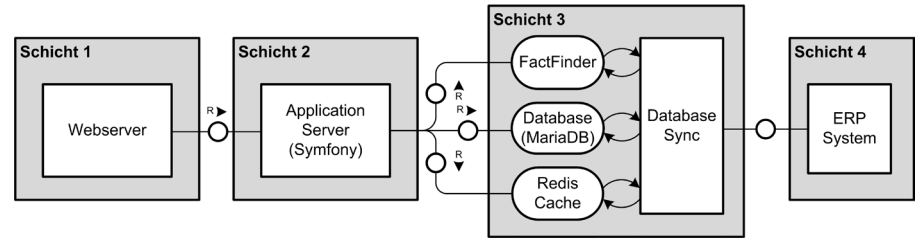

Abbildung 2: FMC-Aufbaustrukturdiagramm der Internetstores-Architektur[18]

zur Verfügung hat. Diese betreffen jedoch nur die Produkt-Detailseite. In der Produktübersicht und der Suche, werden weiterhin die alten Informationen verwendet.[1]

Wie bereits erwähnt, wird für die Suche und Übersicht von Produkten das kommerzielle System Factfinder verwendet. Das Factfinder-System wird lediglich jede Stunde auf den neusten Stand gebracht. Dafür werden die Produkte als CSV-Datei exportiert und vom Factfinder-System wieder importiert. Erst nach diesem Import sind alle Systeme wieder auf dem neusten Stand. Eine Produktänderung kann somit im schlimmsten Fall vom Export des ERP-Systems bis zum Import in Factfinder eine Stunde und 20 Minuten benötigen.[1]

3.4 Anforderungen

An die neue Systemarchitektur werden die folgenden Anforderungen gestellt, welche diese erfüllen sollte, um als Nachfolger für das existierende System in Frage zu kommen.

Produkt-Update-Zyklus
Dadurch, dass eine Produktänderungen über eine Stunde benötigt, bis diese komplett im Shopsystem angekommen ist, kann es zu diversen Problemen kommen. Ein Produkt, das nicht mehr verfügbar ist, weil der Lagerbestand aufgebraucht ist, kann beispielsweise weiterhin bestellt werden. Diese Bestellungen müssen von Mitarbeitern händisch bearbeitet werden, was zusätzliche Kosten verursacht. Der Kunde, dessen Bestellung storniert wird, wird mit der Tatsache, dass er sein gewünschtes Produkt nicht bekommt, nicht zufrieden sein. Die Chance, dass der Kunde erneut etwas kauft, wird dadurch gesenkt. Das gleiche Szenario ist denkbar für ein Produkt, bei dem der Preis nicht korrekt eingetragen ist, wodurch Bestellungen entstehen, die eventuell einen Schaden für das Unternehmen verursachen, da das Produkt zu günstig verkauft wurde. Andersherum werden Produkte, die neu im Sortiment sind, erst nach über einer Stunde im Shop angezeigt und sind damit erst dann zum Kauf verfügbar. Die neue Systemarchitektur muss somit die Zeitspanne, die eine Produktänderung vom ERP-System bis in das Shopsystem benötigt, massiv senken. Konkret sollen die Änderungen, sobald diese verfügbar sind, an das Shopsystem übermittelt werden.[1]

Entkopplung von ERP-System und Shopsystem
Wenn nun die Änderungen an Produkten kontinuierlich an das Shopsystem übertragen werden, kann dadurch die Verfügbarkeit des Shops leiden. Andersherum darf die Last auf dem Shopsystem, welche zu Spitzenzeiten entsteht, die Verfügbarkeit des ERP-Systems nicht beeinträchtigen. Daher muss die neue Architektur die Fähigkeit besitzen, die beiden System voneinander zu entkoppeln.[1]

Schnelle Suche
Die Factfinder-Suche ist im jetzigen System durchaus leistungsfähig. Dies kommt jedoch nur daher, dass Sie in Zusammenarbeit mit

dem Redis-Cache arbeitet. Daher muss die neuen Systemarchitektur auch bei großen Datenmengen eine leistungsfähige Suche anbieten können. Da mit steigenden Zugriffen ebenfalls die Last auf das System zunimmt, muss diese Belastung durch das System ausgeglichen werden können.[1]

Filterinhalte
Wie bereits erwähnt, werden im aktuellen System die Filteroptionen aus den Produktinformationen generiert. Dies hat den Vorteil, dass die Filteroptionen für die Kategorie-Seiten (Produktübersicht) nicht separat gepflegt werden müssen. Die Filteroptionen sind immer auf dem Stand der aktuellen Produktinformationen und zeigen somit das bestmögliche Ergebnis an. Nach der Umstellung der Architektur soll dieser Prozess auch weiterhin möglich sein, sodass auch weiterhin nur die Produktinformationen verwaltet werden müssen.[1]

Austauschbare Datenquelle
Die Datenquelle, die die Produktinformationen bereitstellt, soll austauschbar sein. So kann schneller auf neue Datenbanktechnologien reagiert oder im Falle einer Störung auf ein Backup-System umgestellt werden.[1]

4 Abstrakte Systemarchitektur

Als Grundlage für die neue Architektur wird eine Drei-Schichten-Architektur verwendet. Die Webserver-Schicht sowie die Application-Server-Schicht können unverändert übernommen werden, da hier keine Anpassung der abstrakten Systemkomponenten nötig ist. Lediglich die Datenbank-Schicht wird speziell für die bessere Suche modifiziert.[2]

4.1 Datenbankschicht

Oft werden spezielle Such-Engines verwendet, um einen Index über die vorhandenen Daten zu erstellen und somit eine schnelle Suche zu gewährleisten.[20][19][4][7][17] Daher wird die Datenbank-Schicht um einen Suchindex erweitert, welcher vom Application-Server angesprochen werden kann, sodass dieser anhand der gefundenen Informationen die Daten aus der Datenbank abfragen kann.[15] Daher muss auch weiterhin ein Datenspeicher in der dritten Schicht vorhanden sein, welcher die eigentlichen Produktinformationen vorhält. Dabei muss es sich nicht zwingend um eine relationale Datenbank handeln, da vom Application-Server keine Schreibzugriffe auf Produktdaten stattfinden, sondern nur Lesezugriffe.

Als Alternative zur relationalen Datenbank, welche bisher eingesetzt wurde, kann eine NoSQL-Lösung eingesetzt werden. Eine solche Datenbank ist dafür ausgelegt, eine schnelle Antwortzeit für Lese- und Schreibzugriffe zu haben. Dies ist für eine Produktsuche bzw. für die Abfrage von Produktdaten ideal. Relationale Datenbanken haben eine gewisse Komplexität durch die Verwendung von Transaktionen und ähnlichen Funktionen und sind daher gegenüber NoSQL-Datenbanken im Lese- und Schreibvorgang langsamer. Außerdem sind NoSQL-Datenbanken dafür ausgelegt, sehr große Mengen an Daten zu speichern. Bei der wachsenden Datenmenge ist dies eine notwendige Eigenschaft. Eine relationale Datenbank wird mit steigender Datengröße langsamer und die Antwortzeit erhöht sich. Darüber hinaus eignen sich diese Datenbanken hervorragend für den Betrieb in einem verteilten System (Cluster). NoSQL-Datenbanken haben die Möglichkeit, ohne größere Konfigurationen, verteilt auf mehreren Systemen zu laufen, wodurch die Verfügbarkeit erhöht wird. Zusätzlich kann so theoretisch beliebig horizontal skaliert werden, was ein klarer Vorteil gegenüber relationalen Datenbanken ist.[8][16]

Jedoch haben NoSQL-Datenbanken auch Nachteile. Das sogenannte CAP-Theorem besagt, dass ein Datenbank-System von den drei Eigenschaften Konsistenz (C), Verfügbarkeit (A) und Partitionstoleranz (P) immer nur zwei erfüllen kann. Eine herkömmliche relationale Datenbank ist meist ein CA-System. Ein NoSQL-Cluster ist dagegen partitionstolerant, da der Cluster in zwei Hälften getrennt werden kann und trotzdem noch arbeitet. Wenn ein NoSQL-System ein AP-System ist, ist es für den Benutzer stets verfügbar, selbst wenn es in zwei Hälften geteilt ist. Die Schreibzugriffe sind dann jedoch nicht mehr konsistent. Wenn es sich um ein CP-System handelt, ist es bei einer Aufteilung nicht mehr verfügbar, jedoch wird die Konsistenz gewahrt. Wie das System sich verhalten soll, muss vom Management entschieden werden. Dabei muss die Frage geklärt werden, ob das System immer verfügbar sein soll oder ob Inkonsistenzen wie z.B. eine doppelte Bestellung eines Produkts, das nur noch einmal verfügbar ist, vermieden werden sollen.[14]

Die Datenbank-Schicht kann in diesem Fall auch als Cache für die eigentliche Datenbank gesehen werden, welche sich im ERP-System befindet.[15][6] Diese Arbeit befasst sich zwar nur mit der konkreten Produktsuche, jedoch ist in der allgemeinen Architektur auch weiterhin eine relationale Datenbank vorhanden, um Bestellungen und andere Schreibzugriffe von Benutzern entgegen zu nehmen.

4.2 ERP-Anbindung

Befüllt wird die Datenbank durch das ERP-System, welches entweder als Teil der Datenbankschicht oder als neue Schicht hinter der Datenbankschicht betrachtet werden kann. Das ERP-System speichert die Daten in einer relationalen Datenbank, womit diese auch alle ihre Vorteile genießt und somit vor Dateninkonsistenzen geschützt sind. Die ERP-Schicht bietet eine Schnittstelle, die Datenänderungen für die Datenbank-Schicht bereit stellt. Um die Schichten voneinander zu entkoppeln, soll eine Message-Queue eingesetzt werden, über die die Änderungen transportiert werden. Dabei wird auf das Publish-Subscribe-Verfahren zurückgegriffen werden, bei dem das ERP-System als Publisher dient und die Datenbank-Schnittstelle als Subscriber. Das ERP-System kann somit die Änderungen an Produkten als neue Nachricht an die Datenbank-Schnittstelle senden. Die Schnittstelle wendet die Änderungen auf die Produktinformationen an. So kann nahezu eine Echtzeit-Übertragung erreicht und trotzdem auf die Belastung der verschiedenen Systeme reagiert werden. Die Datenbank-Schicht muss immer nur so viele Nachrichtenpakete abfragen, wie sie verarbeiten kann.[13][12][11][9]

4.3 Vier-Schichten-Architektur

Somit entsteht eine Vier-Schichten-Architektur, wenn man die ERP-Schnittstelle als weitere Schicht sieht. Das ERP-System für sich stellt auch eine Drei-Schichten-Architektur dar, diese ist für die konkrete Produktsuche jedoch nicht relevant. Nur die ERP-Schnittstelle, über die die Produktdaten zur Verfügung gestellt werden, ist Teil der neuen Architektur. Abbildung 3 zeigt die abstrakte vorgeschlagene Architektur, welche als Grundlage für die konkrete Umsetzung verwendet wird. Durch die Abstraktion des ERP-Systems in eine Schnittstelle, können innerhalb des ERP-Systems beliebig Veränderungen vorgenommen werden, ohne dass dies die Shop-Architektur direkt beeinflusst. Dies gilt natürlich nur so lange die Schnittstelle gleich bleibt bzw. die Nachrichten, die über die Message-Queue übermittelt werden, von der Datenbank-Schnittstelle verarbeitet werden können.

5 Umsetzung der Architektur

5.1 Suchindex

Die Such-Engine Lucene gilt als Defakto-Standard in Unternehmen[4], womit sie ein geeigneter Kandidat ist, um den Suchindex für die Produktsuche zu gewährleisten. Jedoch wird auch weiterhin eine Datenbank notwendig sein, um die Informationen der Produkte bereitstellen zu können, wie es in der allgemeinen Systemarchitektur

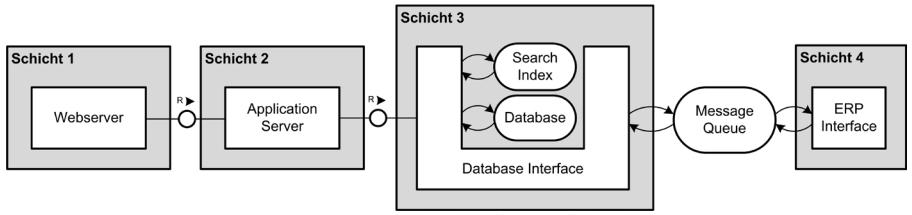

Abbildung 3: FMC-Aufbaustrukturdiagramm der vorgeschlagenen Systemarchitektur[18]

vorgesehen ist. Daher muss beim Einsatz von Lucene ein zusätzlicher Mechanismus entworfen werden, welcher dafür sorgt, dass immer die aktuellsten Informationen aus der Datenbank im Suchindex verfügbar sind.[20][10]

Um diesen Prozess zu vereinfachen, kann eine Datenbank-Lösung gewählt werden, welche Lucene bereits integriert hat. Elasticsearch ist eine solche Datenbank. Neben der effizienten Suche über Lucene bietet Elasticsearch auch eine dokumentenbasierte Speicherung der Daten an. Somit werden der Suchindex und die Datenbank aus der allgemeinen Systemarchitektur zu einer Komponente zusammengefasst. Dadurch werden Dateninkonsistenzen zwischen diesen beiden Systemteilen ausgeschlossen. Ein weiterer Vorteil von Elasticsearch ist seine Eigenschaft als verteiltes System ausgelegt zu sein. Ohne aufwendige Konfiguration kann die Datenbank im Cluster betrieben werden.[4][5]

Die Informationen der einzelnen Produkte werden als JSON-Objekte in der Datenbank abgelegt. Diese werden automatisch in Lucene indiziert. Ein Elasticsearch-Cluster besteht aus mehreren Knoten. Jeder Knoten besitzt eine eigenständige Instanz von Lucene, in der die Informationen, die der Knoten zur Verfügung stellt, indiziert sind. Ein Elasticsearch-Knoten besteht aus sogenannten Shards. Diese repräsentieren eine logische Einheit, welche dazu verwendet wird, die Daten im Cluster zu verteilen. Shards werden in primäre und replizierte Shards unterteilt. Replizierte Shards sind Kopien von primären Shards, um die Redundanz zu erhöhen und somit die Ausfallsicherheit zu steigern. Die Lesezugriffe können von allen Shards beantwortet werden, die Schreibzugriffe nur von den primären. Nach einem erfolgreichen Schreibzugriff, werden die Änderungen in die replizierten Shards übernommen.[5]

5.2 *Produkt-Synchronisation*

Die Synchronisation der Produktdaten wird durch ein Messaging-System gewährleistet, da dadurch eine Entkoppelung der Komponenten stattfindet. Hierfür wird das System RabbitMQ verwendet, da dieses ebenfalls im Cluster betrieben werden kann.[12] Somit kann auch hier die Ausfallsicherheit erhöht wird. Das ERP-System veröffentlicht die Änderungen an Produkten als Nachricht, welche von der Datenbank-Schnittstelle empfangen und als Änderung an Elasticsearch weitergereicht wird. Die Nachricht selbst ist das JSON-Objekt, das in Elasticsearch gespeichert werden soll. Somit muss die Datenbank-Schnittstelle die Informationen nicht weiter verändern, sondern kann diese direkt nach dem Erhalt weiterreichen.[5]

6 Evaluation

6.1 *Anforderungen*

Bei der Analyse der Systemarchitektur wurden diverse Anforderungen aufgestellt, welche eine neue Architektur erfüllen muss. Nachfolgend wird beschrieben, wie die verschiedenen Anforderungen umgesetzt wurden.

Produkt-Update-Zyklus
Der Produkt-Update-Zyklus hat im ursprünglichen System eine Verzögerung von bis zu einer Stunde und 20 Minuten. Dies ist vor allem den verschiedenen Import-Export-Programmen zuzuschreiben. Besonders der verwendete Suchindex Factfinder hat die Verzögerung stark erhöht. In der neuen Architektur wird ein Messaging-System eingesetzt, welches die Änderungen aus dem ERP-System an das Shopsystem überträgt, sobald sie vom ERP-System veröffentlicht wurden. Dadurch erreicht das System einen Update-Zyklus von nahezu Echtzeit. Die einzigen Verzögerungen, die entstehen können, beruhen auf einer extremen Belastung des Shopsystems.

Entkopplung von ERP-System und Shopsystem
Da sowohl ERP-System, als auch Shopsystem unter Belastung durch Benutzerzugriffe stehen, sollten beide Systeme voneinander entkoppelt werden, um diese Last nicht an das jeweils andere System zu übertragen. Durch die Verwendung eines Messaging-Systems wird genau das erreicht. Beide Systeme empfangen bzw. versenden die Pakete so schnell, wie es ihre aktuellen Systemlast zulässt.

Schnelle Suche
In der neuen Architektur wird ein Suchindex verwendet, welcher seine Daten direkt aus dem Datenspeicher des Shops bezieht und nicht auf zusätzliche Cache-Systeme angewiesen ist. In der konkreten Implementierung werden Suchindex und Datenspeicher als eine Einheit umgesetzt.

Filterinhalte
Die Filter, welche in der Produktübersicht angezeigt werden, werden im ursprünglichen System allein durch die Informationen der Produkte befüllt. Dies bleibt auch weiterhin so, um Dateninkonsistenzen zu vermeiden. Die Filteroptionen zeigen immer die Inhalte, welche von den Produktdaten zu Verfügung gestellt werden.

Austauschbare Datenquelle
Die Datenquelle, die das Shopsystem verwendet, ist in der neuen Architektur so implementiert, dass sie durch eine beliebig andere ersetzt werden kann. Falls eine neue Datenquelle verwendet werden soll, müssen lediglich die entsprechenden Schnittstellen umgesetzt und im Shopsystem auf die neue Datenquelle umgestellt werden. Dies ermöglicht es, ohne größeren Eingriff in das Shopsystem, schnell auf neue Datenspeicherlösungen reagieren zu können. Zusätzlich bietet es die Möglichkeit Fallback-Lösungen zu implementieren.

6.2 *Leistungsvergleich*

Um die konkrete Leistungssteigerung in Zahlen zu messen, wurden das alte System mit Factfinder und das neue System mit Elasticsearch verglichen. Dazu wurden zwei Leistungstests durchgeführt, welche jeweils mit aktivem Cache und inaktiven Cache im Shopsystem gemessen wurden. Der Redis-Cache ist bei der Messung des alten Systems aktiv geblieben, da Factfinder diesen benötigt, um die Produktinformationen abrufen zu können. Jeder Test wurde jeweils 50 mal durchgeführt. Das Ergebnis ist der Mittelwert aus allen Messungen.

6.3 *Testfall 1*

Der erste Test misst den Zeitraum, ab dem die Anfrage für eine Produktübersicht an die Datenbank gesendet wurde bis zum Zeitpunkt der Antwort, inklusive der Erzeugung der Produkt-Objekte. Dabei sind bei aktivem Cache folgende Messergebnisse entstanden:

Tabelle 1: Testfall 1 mit Cache

System	Messwert
Alt	15ms
Neu	8ms

Der gleiche Testfall mit inaktivem Cache ergab das nachfolgende Ergebnis:

Tabelle 2: Testfall 1 ohne Cache

System	Messwert
Alt	10.194ms (10s)
Neu	2.443ms (2s)

6.4 Testfall 2

Der zweite Test misst den gesamten Seitenaufruf, bei dem die Seite komplett geladen wird. Dabei sind bei aktivem Cache folgende Messergebnisse entstanden:

Tabelle 3: Testfall 2 mit Cache

System	Messwert
Alt	842ms
Neu	571ms

Der gleiche Testfall mit inaktivem Cache, ergab folgende Messwerte:

Tabelle 4: Testfall 2 ohne Cache

System	Messwert
Alt	14.043ms (14s)
Neu	4.124ms (4s)

6.5 Ergebnis

In allen Testfällen ist die neue Systemarchitektur mit der prototypischen Implementierung schneller, als das alte System. Im ersten Testfall ist das neue System um 47% (mit Cache) bzw. 76% (ohne Cache) schneller, im zweiten Testfall ist es um 32% (mit Cache) bzw. 71% (ohne Cache) schneller. In allen Testfällen wurde eine kontinuierliche Last durch andere Benutzer auf das System simuliert. Dabei waren immer ca. 100 weitere Benutzer auf der Seite aktiv.

7 Fazit

Die Analyse des alten Systems zeigt extreme Defizite im Update-Zyklus der Produktdaten. Diese entstehen durch die komplexe Datenbankstruktur und die verschiedenen Import- und Export-Programme, welche nicht aufeinander abgestimmt sind.

Die vorgeschlagene Architektur überarbeitet das Systemmodell so, dass die Änderungen kontinuierlich in die Datenbank eingepflegt werden. So werden Änderungen so schnell wie möglich im Shopsystem sichtbar. Um die Produktsuche zu verbessern, wird neben einem Datenspeicher auch ein spezieller Suchindex eingeführt. Dadurch können Suchanfragen schnell beantwortet und die entsprechenden Informationen aus dem Datenspeicher abgerufen werden.

In der konkreten Lösung werden der Datenspeicher und der Suchindex als eine Softwarelösung abgebildet. Elasticsearch ist eine dokumentenbasierte Datenbank mit integriertem Suchindex. Durch die Verwendung einer Lösung, die beide Elemente vereint, muss nicht darauf geachtet werden, dass der Suchindex immer die aktuellsten Informationen aus dem Datenspeicher indiziert hat.

Um die Informationen aus dem ERP-System in das Shopsystem zu übertragen, wurde bisher eine Import- Export-Funktion mit XML-Dateien verwendet. Die neue Architektur entkoppelt die beiden Systeme durch ein Messaging-System. Dadurch wird die Last, die auf den beiden Systemen liegt, nicht auf das jeweils andere übertragen. Außerdem ermöglicht dies die Einführung beliebig weiterer Datenquellen, die beispielsweise als Fallback-Lösung oder als Alternative zu Elasticsearch verwendet werden können. Das ERP-System muss nicht jedes Mal angepasst werden, wenn im Shopsystem die Datenquelle verändert oder angepasst wird.

Ein weiterer Aspekt, der in dieser Arbeit nicht behandelt wurde, jedoch die Leistung des Shopsystems massiv steigern würde, ist die Überarbeitung der Shop-Implementierung. Das Shopsystem ist historisch gewachsen und weist daher diverse Schwächen auf. Das Produkt-Objekt beispielsweise, welches vom System verwendet wird, um die Produkt-Informationen zu speichern, ist in der Implementierung kein reines Datenobjekt mehr, sondern besitzt auch aktive Komponenten. Dadurch erhöht sich die Ladezeit einer Instanz von dieser Klasse enorm. Grundsätzlich sollte das Shopssystem mehr Hilfsmittel des PHP-Frameworks Symfony verwenden, auf dem das Shopsystem aufbaut.

Literatur

[1] Internetstores gmbh, 2015.

[2] D. P. S. Andrew Boyer, Bernd Bruegge. Evaluating e-commerce cluster architectures using simulation. *E-Commerce Technology, 2005. CEC 2005. Seventh IEEE International Conference*, 1:135 – 142, 2005.

[3] D. A. G. Anuradha S. Kanade. Choosing right database system: Row or column-store. *Information Communication and Embedded Systems (ICICES), 2013 International Conference*, 1:16 – 20, 2013.

[4] J. Bai. Feasibility analysis of big log data real time search based on hbase and elasticsearch. *Natural Computation (ICNC), 2013 Ninth International Conference*, 1:1166 – 1170, 2013.

[5] Z. T. Clinton Gormley. *Elasticsearch: The Definitive Guide*. O'Reilly Media, Sebastopol, 2015.

[6] N. R. Felix Gessert, Florian Bücklers. Orestes: A scalable database-as-a-service architecture for low latency. *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference*, 1:215 – 222, 2014.

[7] J. L. Hongbin Zhang. Search engine design based on web service and lucene. *Information Engineering, 2009. ICIE '09. WASE International Conference*, 1:458 – 461, 2009.

[8] G. L. J. D. Jing Han, Haihong E. Survey on nosql database. *Pervasive Computing and Applications (ICPCA), 2011 6th International Conference*, 1:363 – 366, 2011.

[9] R. G. Kangseok Kim, Marlon E. Pierce. Sqmd: Architecture for scalable, distributed database system built on virtual private servers. *eScience, 2008. eScience '08. IEEE Fourth International Conference*, 1:658 – 665, 2008.

[10] U. H. Luiz André Barroso, Jeffrey Dean. Web search for a planet: The google cluster architecture. *IEEE Computer Society 2003*, 1:22 – 28, 2003.

[11] A. K. D. K. Markus Keidl, Alexander Kreutz. A publish and subscribe architecture for distributed metadata management. *Data Engineering, 2002. Proceedings. 18th International Conference*, 18:309 – 320, 2002.

[12] J. B. R. G. J. N. Nicolas Viennot, Mathias Lécuyer. Synapse: a microservices architecture for heterogeneous-database web applications. *Proceedings of the Tenth European Conference on Computer Systems Article No. 21*, 21, 2015.

[13] R. G. A.-M. K. Patrick Th. Eugster, Pascal A. Felber. The many faces of publish/subscribe. *ACM Computing Surveys (CSUR)*, 1:114 – 131, 2003.

[14] M. F. Pramod J. Sadalage. *NoSQL Distilled - A Brief Guide to the Emerging World of Polyglot Persistence*. Pearson Education, Inc., New Jersey, 2013.

[15] J. W. Qi Su. Indexing relational database content offline for efficient keyword-based search. *Database Engineering and Application Symposium, 2005. IDEAS 2005. 9th International*, 1:297 – 306, 2005.

[16] R. D. Richard K. Lomotey. Terms mining in document-based nosql: Response to unstructured data. *Big Data*

(BigData Congress), 2014 IEEE International Congress, 1:661 – 668, 2014.

[17] W.-T. B. Silviu Homoceanu. Querying concepts in product data by means of query expansion. *Web Intelligence and Agent Systems: An International Journal 12 (2014)*, 1:1 – 14, 2014.

[18] P. Tabeling. *Softwaresysteme und ihre Modellierung*. Springer-Verlag, Berlin Heidelberg, 2006.

[19] O. K. e. a. Thorsten Schlachter, Clemens Düpmeier. Towards a search driven system architecture for environmental information portals. *IFIP International Federation for Information Processing 2015*, 1:351 – 360, 2015.

[20] Z. W. Xiujin Shi. An optimized full-text retrieval system based on lucene in oracle database. *Enterprise Systems Conference (ES), 2014*, 1:61 – 65, 2014.

Entwicklung gestenbasierter Zeigerinteraktionen für Augmented Reality Anwendungen

Florian Strieg
Reutlingen University
Florian.Strieg@Student.
Reutlingen-University.DE

Abstract

In dieser Arbeit wird ein Plug-In vorgestellt, welches Sensordaten, sowie die erfassten Gesten der Thalmic Myo für Androidgeräte unter Verwendung der Unity Game-Engine bereitstellt. Es wurden weiterhin drei gestenbasierte Interaktionsmethoden implementiert, die das Steuern eines Zeigers in Augmented Reality Anwendungen ermöglicht. Um erste Rückmeldungen für eine mögliche Weiterentwicklung der Zeigemethoden zu erhalten, wurde eine Studie mit dem Fragebogen aus ISO 9241-9 und der Think-Aloud Methode durchgeführt. Die Ergebnisse zeigen, dass sich die Interaktionsmethoden bei der Erlernbarkeit und Genauigkeit nur wenig unterscheiden, eine der Methoden aber einen erheblich höheren Kraftaufwand bei der Verwendung benötigt.

Schlüsselwörter

Erweiterte Realität, GearVR, Thalmic Myo, Gestensteuerung, Zeigeinteraktion

---

Betreuer Hochschule: Prof. Dr. Uwe Kloos
Hochschule Reutlingen
Uwe.Kloos@Reutlingen-University.de
Betreuer Firma: Dr. Matthias Bues
Fraunhofer IAO
Matthias.Bues@iao.fraunhofer.de

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

CR-Kategorien

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—artificial, augmented, and virtual realities; H.5.2 [Information Interfaces and Presentation]: User Interfaces–Input devices and strategies (e.g., mouse, touchscreen)

1 Einleitung

Das Forschungsprojekt „Glassroom“ beschäftigt sich mit der Umsetzung von Bildungskonzepten unter Verwendung virtueller und erweiterter Realität (VR und AR). Der Fokus liegt auf der beruflichen Aus- und Weiterbildung von Fachkräften im Maschinen- und Anlagenbau. Vorangegangene Arbeiten aus diesem Projekt arbeiteten an reinen VR-Anwendungen in denen Benutzer mithilfe von Head-Mounted-Displays (HMDs) Montageaufgaben bearbeiten.

Forschungen haben bereits gezeigt, dass die eigene Wahrnehmung in virtuellen Umgebungen durch viele Faktoren beeinflusst werden können. Beispielsweise macht das Fehlen von realen Bezugspunkten eine Einschätzung von Distanzen und Größen schwierig [1]. Zusätzlich benötigt der Übergang in eine komplett virtuelle Welt eine Eingewöhnungsphase, die in VR-Anwendungen mit Modellierungen des realen Testraums umgangen wird. Beide Methoden zur Steigerung der Immersion sind mit einem gesteigerten technischen Aufwand verbunden [1]

Um diese Aspekte auszuklammern, soll in einem zukünftigen Projekt untersucht werden, wie sich eine kamerabasierte Umgebungssicht gegenüber einer reinen VR Umgebung auswirkt. Da das Kamerabild direkt auf das Display des Anwenders übertragen wird, wird diese Technik auch Video-See-Through genannt.

Im Vorfeld sollen in dieser Arbeit mögliche Interaktionsmethoden, speziell für das Steuern eines Zeigers für derartige Umgebungen, vorgestellt werden. Hierfür werden zwei gestenbasierte und eine blickbasierte (Gaze Tracking) Zeigemethode implementiert. Um einen Prototyp auch für spätere Einsatzzwecke so einfach und kostengünstig wie möglich zu halten, werden VR- und AR-Geräte verwendet, die aus dem Consumermarkt stammen. Als HMD kommt die erste Version der GearVR von Samsung, in Verbindung mit dem Galaxy Note 4, zum Einsatz. Um Gesten des Anwenders zu erfassen, wird das Armband Myo von Thalmic verwendet.

2 Verwandte Arbeiten

In einer Arbeit von Haque, Nancel und Vogel [2] verwenden die Autoren das Myo Armband, um eine freihändige Zeigerinteraktion auf großen Displays und Powerwalls zu ermöglichen. Eine Evaluierung bestand aus einem Fitts Law Test, der die Genauigkeit und Geschwindigkeit ihrer Anwendung gegenüber einem Kamerabasierten Tracking verglich. Ihre Testergebnisse zeigen, dass Myo nur geringe Unterschiede in den untersuchten Eigenschaften aufweist. Besonders interessant sind auch, die verschiedenen Korrektur- und Filterfunktionen die sie in ihrem System implementiert haben, um Schwächen in der Verwendung mit der Myo auszugleichen. Als mögliche zukünftige Arbeit schreiben die Autoren über die Verwendung der Myo in Zusammenhang mit Smartphones und HMDs, was in dieser Arbeit aufgegriffen werden soll. [2]

In 2003 führten Cournia, Smith und Duchowski [3] einen Vergleich zwischen Gaze- und Gesten-Zeigeinteraktionen in einer virtuellen Umgebung durch. Probanden sahen über ein HMD 20 verschiedene Objekte an zufällig platzierten Positionen im dreidimensionalen Raum und mussten diese per Augenbewegung und Handbewegung anwählen. Nach Auswertung ihrer Ergebnisse konnten die Autoren bei naher und mittlerer Distanz der Objekte keinen klaren Geschwindigkeitsunterschied zwischen den beiden Methoden erkennen. Nur bei weiter entfernten Objekten war die Gaze-Methode leicht im Vorteil. [3]

Douglas, Kirkpatrick und MacKanzie evaluieren in ihrer Arbeit „*Testing Pointing Device Performance and User Assessment with the IS0 9241, Part 9 Standard* “ ob der ISO Standard für Zeigegeräte wissenschaftlich verwertbare Ergebnisse liefert [4]. Ein wichtiger Punkt ihrer Ergebnisse war das Abweichen von einer 7 Punkte auf eine 5 Punkte Skala im Fragebogen, da ihre Probanden keine feinere Unterscheidung zu den Fragen machen konnten. Weiterhin empfanden sie die Ergebnisse des Tests alleine als zu vage, weswegen sie ein zusätzliches Interview am Ende der eigentlichen Fragebögen vorschlagen [4].

3 Implementierung

Für den Prototyp der Zeigermethoden soll Unity als Entwicklungsumgebung verwendet werden. Ein erstes Hindernis ist hier die Anbindung der Myo über Unity an das Androidgerät Galaxy Note 4. In diesem Kapitel soll die verwendete Hard- und Software, die Entwicklung des Unity Plug-Ins und der Zeigemethoden vorgestellt werden.

3.1 Materialien

Für die Entwicklung der Anwendung wurde Unity auf einem Windows 7 64bit Betriebssystem verwendet. Das Galaxy Note 4 ist mit einem Snapdragon 805 2,7 GHz QuadCore Prozesser, 3GB RAM und einem 5,7 Zoll Display mit einer Auflösung von 1440 x 2560 Pixel ausgestattet. Die Auflösung wird von Unity in zwei 1024 x 1024 Texturen aufgeteilt. Zusammen mit dem GearVR Headset kann die Hardware als HMD verwendet werden und ermöglicht das Abspielen von VR- und AR-Anwendungen. Zusätzlich können die Rotationsdaten des Gerätes, die dann von dem Kopf des Anwenders manipuliert werden, verwendet werden. Für die Aufnahme der Umgebung bietet das Smartphone eine 16 Megapixel Kamera, weswegen das Bild für den Anwender nur monoskopisch dargestellt werden kann. Um den Video-See-Through auf der Softwareseite zu ermöglichen, wurde die AR-Plattform Vuforia eingesetzt. Diese unterstütz sowohl Unity in der 32bit Version, als auch das GearVR. Um virtuelle Objekte im Raum zu positionieren, nutzt die Plattform Bildverarbeitung, um Bilder oder einfache reale Objekte zu erkennen. Da der fertige Prototyp auf dem Galaxy Note 4 läuft, waren die Ansprüche an den PC auf dem entwickelt wurde dementsprechend niedrig, weswegen ein Notebook mit 4GB Arbeitsspeicher und einem Intel Core 2 Duo 2.13GHz P7450 ausreichte.

Das Myo Armband von Thalmic integriert EMG und IMU Sensoren und kann somit seine Rotation in drei Freiheitsgraden, sowie ausgeführte Gesten erkennen. Die Gesten, die erkannt werden können, beschränken sich auf fünf vordefinierte Bewegungen mit der Hand. Unter Zuhilfenahme der Rotation können jedoch weitere Gesten selbst definiert werden. Die IMU-Daten können verwendet werden, um die Rotation der Myo zu bestimmen, nicht jedoch die Position im Raum. Dies hat zur Folge, dass zumindest eine Zeigeinteraktion im virtuellen Raum auf eine zweidimensionale Ebene beschränkt ist. Für die Übertragung der Daten wird der Bluetooth 4.0LE Adapter des Smartphones verwendet.

3.2 Plug-In

Für die Entwicklung von Anwendungen für das Gestenarmband Myo stellt Thalmic mehrere Softwareentwicklungskits zur Verfügung. Unter anderen finden sich SDKs für die Entwicklung mit Android Studio und für die Game-Engine Unity. Leider gibt es mit den offiziellen Werkzeugen nicht die Möglichkeit Anwendungen, die die Myo als Interaktionsgerät verwenden und in Unity geschrieben, wurden auch auf eine Android-Plattform auszuliefern. Da dies aber eine Grundvoraussetzung für unseren Prototyp ist, musste zunächst die Schnittstelle zwischen diesen Plattformen geschrieben werden.

Hierzu wurde, mithilfe des Myo Android SDKs und Android Studio, ein Plug-In entwickelt, welches in Unity verwendet werden kann. Wird die Anwendung dann von Unity für ein Androidgerät gebuildet, wird das Plug-In mitgeliefert. Eines der Hauptprobleme hierbei war das Anwendungsmanagement des Android Systems. Weil Vuforia im Android-Manifest bereits den Platz für die Activity belegt, konnte das Myo Plug-In nicht von dort aufgerufen werden. Um dies zu umgehen, wurde das Plug-In zu einem Service umgeschrieben. Dieser hält sich nun im Hintergrund der eigentlichen Anwendung und beliefert diese nur mit Statusupdates des Armbands wenn diese eintreten.

Um den Entwicklungsablauf zu vereinfachen, wurde die Lösung auch noch in ein bereits vorhandenes Plug-In eingebaut, welches Windows und iOS unterstützt. Damit ist es also möglich, das Armband während der Entwicklung am PC zu verwenden, um seine Anwendung direkt im Editor zu testen und im Anschluss daran sowohl für Android als auch für iOS zu builden. Das Plug-In und dessen Quellcode wurden zur freien Verfügung auf GitHub veröffentlicht.[1]

[1] https://github.com/f-strieg

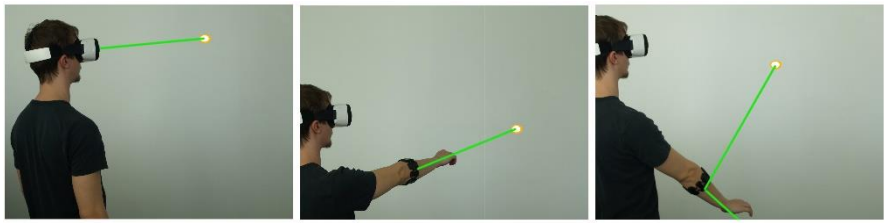

Abbildung 1: Veranschaulichte Zeigemethoden Gaze, Direkt und Indirekt (v.l.n.r.)

3.3 Zeigemethoden

Für den Prototyp wurden drei Interaktionsmethoden implementiert. Um diese einfacher zu unterscheiden, wurden sie Gaze, Direkt und Indirekt genannt. Mithilfe aller drei Methoden kann der Anwender auf der Ebene, auf der das Videobild angezeigt wird, einen Zeiger bewegen. Dieser wird durch einen gelben Kreis symbolisiert. Als Feedback wann sich der Zeiger und ein virtuelles Objekt sich überlappen und dieses dann angeklickt werden kann, verfärbt sich der Zeiger grün. Der Zeiger befindet sich immer direkt vor der Videoebene. Damit der Zeiger trotzdem nicht hinter Objekten verschwindet, die sich zwischen dem Video und dem Anwender befinden, wird ein Overlay-Shader verwendet. So erscheint der Zeiger immer im Vordergrund. Um zu erkennen, wann sich ein Objekt vor dem Zeiger befindet, wird ein weiterer Raycast in die Vorwärtsrichtung des Zeigers verwendet.

Für das Auslösen eines Klicks wurde bei allen drei Zeigemethoden die Spreizgeste der Myo verwendet. Dabei spreizt der Anwender die Finger auseinander. Um sicher zu gehen, dass die Geste richtig erkannt wurde, wird als Feedback ein kuzer Piepton abgespielt. Dies bedeutet allerdings nicht, dass das Objekt auch getroffen wurde. Da sich bei der Ausführung einer Geste der Arm bewegt, kommt es vor, dass der Zeiger von einem angewählten Objekt rutscht, bevor der Klick richtig durchgeführt werden kann. Haque et. al. fanden heraus, dass sich diese Form des fehlerhaften Klicks in den meisten Fällen zwischen 250 und 500ms nach der Ausführung einer Geste ereignet [2]. Um dem entgegenzuwirken wurde eine Korrekturfunktion implementiert, die das zuletzt anvisierte Objekt über ein gewisses Zeitfenster hinweg speichert. In unserem Prototyp wurde dieser Wert auf 300ms festgesetzt. Ein Klick wird dann auf dieses Objekt ausgeführt, auch wenn es sich nicht mehr unter dem Zeiger befindet.

Die Gaze-Methode beruht auf den Rotationsdaten des HMDs. Sie wird schon lange in VR-Anwendungen verwendet [3] und ist derzeit in Apps, die im Google AppStore erhältlich sind, sehr populär [5]. Von der Kameraposition aus wird ein Raycast in Vorwärtsrichtung ausgesendet. Wird eine Kollision mit der Videoebene festgestellt, wird der Zeiger an dieser Stelle für den Anwender eingeblendet. Da sich die Videoebene immer in Blickrichtung des Anwenders befindet ist dies bei dieser Methode immer der Fall.

Die Direkt-Methode verwendet statt der Rotation des HMDs, die des Myo Armbands um den Zeiger zu bewegen. Dabei zeigt der Anwender mit seinem Arm in die Richtung in der sich das Bild des Video-See-Throughs befindet. Der Raycast für die Ermittlung der Kollision geht hier in die Richtung, in die die Myo ausgerichtet wird. Der Ursprung des Raycasts ist bei den beiden gestenbasierten Zeigemethoden der gleiche, wie bei der Gaze-Methode. Es kann also vorkommen, dass die Videoebene bei Verwendung der Myo nicht getroffen wird. Ist dies der Fall, verbleibt der Zeiger an der letzten bekannten Kollisionsposition. Der Raycast wird auch hier zu jeder Zeit ausgesendet. Diese Form des Zeigens wurde auch schon oft untersucht [2][3]. Zusätzlich wird die Rotation des HMDs auf die der Myo aufgerechnet, womit ermöglicht wird, dass der Anwender sich 360° in jeder Achse des dreidimensionalen Raums drehen kann und jeder Zeit die Möglichkeit hat mit seinem Arm zu zeigen.

Auf dieser Grundlage wurde noch eine dritte Alternative, die Indirekt-Methode, implementiert. Aus eigenen Erfahrungen ging hervor, dass es sehr ermüdend ist, den Arm über längere Zeit ausgestreckt vor sich zu halten. In der Indirekt-Methode wird der Raycast der von dem Myo Armband ausgeht nicht in dessen Vorwärtsrichtung ausgesendet, sondern in einem positiven Winkel zur x-Achse. Ist der Blick nach vorne gerichtet beträgt dieser Winkel 45°, wodurch der Zeiger sich in der Mitte des Bildes befindet, wenn der Anwender seinen Arm im selben Winkel nach unten hält. Es wurde zusätzlich eine Funktion implementiert, die den Winkel anpasst, je nachdem ob der Anwender weiter nach unten oder nach oben blickt. Bleibt der Arm in demselben Winkel, verändert sich die Zeigerposition also nicht wenn sich die Rotation des HMDs auf der Querachse ändert. Nach bestem Wissen des Autors ist dies eine neue Herangehensweise unter Verwendung gestenbasierter Zeigeinteraktion.

Bei der Verwendung der Myo muss darüber hinaus beachtet werden, dass die Rotationsdaten des Armbands zu Beginn der Anwendung zur virtuellen Welt ausgerichtet werden müssen. Um dies zu bewerkstelligen wurde die Faustgeste verwendet, die der Anwender einmalig zu Beginn der Anwendung ausführen muss. Diese Kalibrierung richtet die virtuelle Präsentation der Myo im Raum aus und kann wiederholt werden falls die Position des Zeigers nicht mehr die der Myo wiederspiegeln sollte.

4 Evaluation

In diesem Kapitel sollen die Methodik, der Ablauf und die Ergebnisse einer ersten Evaluierung, der drei Zeigemethoden des Prototyps, beschrieben werden. Hierzu wurden ein angepasster Fragebogen aus der ISO Norm 9241-9 (Abbildung 2) für die Evaluation von Zeigegeräten und die Think-Aloud Methode verwendet. Das Ziel dieser Evaluierung war ein Stimmungsbild zu den einzelnen Methoden zu erhalten um frühzeitig entscheiden zu können, ob eine Weiterentwicklung für die einzelnen Interaktionen sinnvoll ist. Besonderer Fokus wurde auf die von Cabral et. Al. vorgebrachten Anforderungen an Human-Computer-Intefaces gelegt. Diese sind eine leichte Erlernbarkeit der Bedienung, sowie die Effizienz und der Komfort der eingesetzten Gesten. [6]

1. Die Steuerung war
   (1: sehr grob – 5: sehr flüßig)
2. Die geistige Anforderung die für die Steuerung benötigt wurde war
   (1: niedrig– 5: hoch)
3. Die physische Anstrengung die für die Steuerung nötig wurde war
   (1: niedrig – 5: hoch)
4. Akkurates ansteuern der Ziele war
   (1: einfach – 5: schwer)
5. Die Anwendungsgeschwindigkeit war
   (1: zu schnell – 5: zu langsam)
6. Müdigkeit der Finger
   (1: keine – 5: sehr hoch)
7. Müdigkeit des Handgelenks
   (1: keine – 5: sehr hoch)
8. Müdigkeit des Arms
   (1: keine – 5: sehr hoch)
9. Müdigkeit der Schulter
   (1: keine – 5: sehr hoch)
10. Müdigkeit des Nacken
    (1: keine – 5: sehr hoch)
11. Der allgemeine Komfort der Steuerungsmethode war
    (1: sehr unangenehm– 5: angenehm)
12. Die Verwendung der Steuerungsmethode war
    (1: sehr schwierig zu bedienen– 5: sehr einfach zu bedienen)

Abbildung 2: Verwendete Fragen aus ISO 9241-9

4.1 Methodik

Um neben den quantitativen Ergebnissen der Fragebögen auch qualitative Antworten zu erhalten, wird in unseren Tests die Think-Aloud Methode verwendet. Bei dieser Methode soll der Proband alle Gedanken, die er bezüglich der von ihm getesteten Anwendung hat, laut aussprechen. Hierbei wird auf eine einfache Art Feedback zum Prototyp gesammelt. Dies kommt dem Vorschlag von Monk, Wright und Haber entgegen, die Fragebögen durch die direkten Aussagen von Nutzern zu erweitern, um die Vagheit der Fragebögen auszugleichen. [7]

Wegen des frühen Stadiums des Prototyps wurde zudem ein kooperativer Ansatz der Tests gewählt. Hierbei kann der Beobachter aktiv in den Ablauf eingreifen, sollte es Probleme mit der Anwendung geben [7]. Nachteile dieser Art der Think-Aloud Methode beschreibt Bevan in "*Usability is quality of use*" [8]. Nach seiner Aussage kann man nach dem aktiven Eingriff in den Testablauf nicht erfahren, ob der Proband Probleme bei der Verwendung in einem realen Umfeld selbst hätte lösen können. Wichtiger noch sind die Aussagen, dass eine akkurate Zeitmessung für die Vollendung der Aufgaben dann nicht mehr möglich ist sowie das Fragen zum Komfort der Anwendung verfälscht werden können. [8]

Ein Fitts Law Test, der ein weiterer Hauptbestandteil der verwendeten ISO Norm ist, wurde für diese Arbeit deswegen vorerst noch nicht vorgesehen. Um die Aussagen der Probanden zum Komfort der Steuerung nicht zu verfälschen, wird nur bei einer Fehlfunktion des Prototyps eingegriffen. Zusätzlich sollen die Ergebnisse des Think-Aloud Tests mit denen der Fragebögen verglichen werden.

Wir gehen davon aus, dass ein kleiner Pool an Probanden, unter Berücksichtigung des Ziels ein erstes Feedback zu erhalten, für diese frühe Iteration des Prototyps ausreicht [7]. Da das Projektumfeld zudem nicht für Experten gedacht ist, wurden 6 Probanden rekrutiert die vorher noch keine Erfahrung mit dem Myo Armband gemacht haben. Teilweise wiesen die Probanden Erfahrung mit VR Anwendungen und HMDs auf, nicht aber mit ähnlichen Methoden des Zeigens im virtuellen oder augmentierten Raum. Unter den Probanden war ein Teilnehmer unter 20 Jahren, der Rest zwischen 21 und 30. Weiterhin nahmen 2 Frauen, 3 Brillenträger und 2 Linkshänder teil.

4.2 Durchführung

Als Aufgabe für die Probanden wurde eine Abwandlung des One-Tapping Tests und im Anschluss des Multi-Tapping Tests durchgeführt. Hierbei müssen die Teilnehmer zunächst 20 virtuelle Boxen, die ihre Größe und Entfernung zueinander ändern, anwählen und die Klickgeste ausführen. Beim Multi-Tapping Test erscheinen vor dem Probanden Kreise, wobei diese in einer festgelegten Reihenfolge, markiert durch eine rote Einfärbung, angeklickt werden müssen.

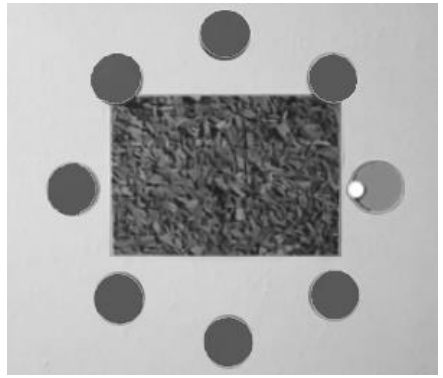

Abbildung 3: Multiple-Tap Test in der AR Anwendung

Für die Darstellung der virtuellen Objekte der Tests wurde ein Trackingtarget auf Kopfhöhe der Probanden in 90cm Entfernung aufgehängt. Um optimale Ergebnisse zu erhalten, wurde jeder Proband vor Beginn durch das Kalibrierungsprogramm von Thalmic geleitet. Dies sollte sicherstellen, dass das Armband die Gesten für jeden Anwender individuell besser erkennt. Den Teilnehmern wurde dann die jeweils erste Zeigemethode erklärt, woraufhin sie die Aufgabe mit dieser durchführten. Im Anschluss daran füllte jeder Teilnehmer den Fragebogen zur Methode aus, bevor ihm die nächste Methode erklärt wurde. Weil der Fragebogen auch nach der Müdigkeit einzelner Körperteile nach der Verwendung der Interaktionsmethode fragt, wurden die 6 Probanden in den 6 möglichen Reihenfolgen durch die Methoden geführt. Damit sollte verhindert werden, dass eine anstrengendere Zeigemethode Seiteneffekte auf die nachfolgende Methode hat. Nachdem der Proband alle drei Methoden getestet hatte wurde er noch gefragt, welche er am besten und welche am schlechtesten fand. Zusätzlich wurde ihm die Möglichkeit gegeben, seine Entscheidung zu begründen und weitere Anmerkungen zur Anwendung zu machen.

5 Ergebnisse und Diskussion

5.1 Fragebogen

Nach der Aufschlüsselung der Antworten aus dem Fragebogen, nach dem arithmetischen Mittel (Abbildung 4) und den einzelnen Methoden, lassen sich nicht in allen Fragen signifikante Unterschiede feststellen. Die Steuerung des Zeigers mit den unterschiedlichen Interaktionsformen wurde von allen Probanden als ähnlich flüssig und einfach eingestuft. Positiv ist das diese Werte recht gut ausgefallen sind. Wie erwartet waren die Werte für die Direkt-Methode bei der Müdigkeitserscheinung des Arms und der Schulter sowie der für die physische Anstrengung sehr hoch und der damit korrelierende Wert des allgemeinen Komforts dementsprechend niedrig. Hier schneiden die beiden anderen Methoden deutlich besser ab.

Wir vermuten, dass der erhöhte Wert für die Indirekt-Methode bei der Müdigkeit der Finger auf die Ungenauigkeit des Fragebogens zurück zu führen ist, da die verwendeten Gesten jeweils dieselben waren. Die ungewohnte Art der Steuerung der Indirekt-Methode ist in den Ergebnissen nicht negativ aufgefallen, schneidet bei der Einfachheit genauso gut ab, wie die Gaze-Methode und hat nach den Probanden den höchsten Komfort bei der Bedienung.

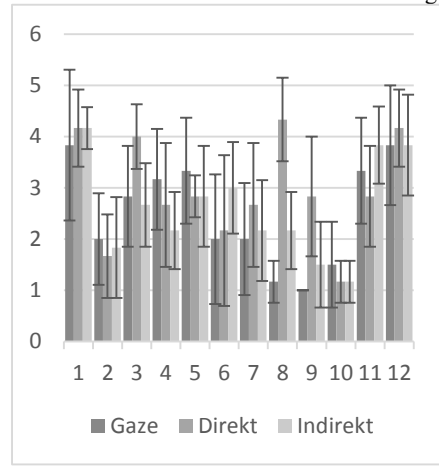

Abbildung 4: Ergebnisse des Fragebogens als arithmetisches Mittel

5.2 Think-Aloud

Die Auswertung der Think-Aloud Methode ergab für die Gaze-Methode, dass vier der Probanden sie als angenehm und sehr akkurat empfanden, was mit den Ergebnissen des Fragebogens übereinstimmt. Alle Probanden stellten hier einen gewissen Lerneffekt fest, der es ihnen ermöglichte die Ziele mit der Zeit schneller zu treffen. Nur zwei der Teilnehmer empfanden es als schwierig die Ziele zu treffen, weil sie oft darüber hinaus schossen. Das Ausführen der Spreizgeste fühlte sich für zwei Probanden im Vergleich zur Direkt-Methode einfacher an, da der Arm nicht angehoben werden muss.

Die Direkt-Methode empfanden durchweg alle Probanden als sehr anstrengend. Zusätzlich dazu schien das Ausführen der Spreizgeste hier anstrengender zu sein. Trotzdem konnten sie die Ziele gut anvisieren und Klicks ausführen. Ein Proband sagte aus, dass er diese Interaktionsmethode am natürlichsten fand.

Obwohl zwei Teilnehmer die Indirekt-Methode als ungewohnt empfanden, wurde diese Interaktion mit dem Zeiger von allen als präzise beschrieben. Ein Teilnehmer sagte, dass es ihm hier leichter fällt die kleineren Ziele zu treffen, als mit den anderen Methoden. Alle Probanden sagten aus, dass die Methode weniger anstrengend ist als die Direkt-Methode. Ein Teilnehmer deckte allerdings ein Problem mit der Methode auf. Lässt man den Arm fast ganz hängen, befindet sich der Zeiger auf dem Bildschirm am unteren Rand. Als Rechtshänder hat man nun das Problem nach Links zu zeigen, da der eigene Körper im Weg ist. Als Linkshänder ist es genau umgekehrt.

Weiteres Feedback betraf auch die Anwendung an sich. Nur ein Proband empfand die Auflösung der GearVR mit dem See-Through als zu gering und bemerkte eine gewisse Latenz. Alle Probanden befanden das gelieferte Feedback der Anwendung als ausreichend. Obwohl versucht wurde die Lichtverhältnisse so konstant wie möglich zu halten, um ein optimales Tracking der Vuforia Targets zu ermöglichen, schwankten diese je nach Tageszeit minimal. Dennoch fiel den Probanden das einblenden der virtuellen Objekte nicht negativ auf, was für eine weitere Verwendung der Vuforia Software spricht.

Nach den Aussagen der Probanden ist die Spreizgeste zum Ausführen der Klicks eher ungeeignet. Beachtet werden muss hier allerdings, dass die Probanden keine Erfahrung mit der Myo hatten. Aus eigener Erfahrung fällt das Ausführen der Gesten mit der Zeit leichter, da Anwender dazu neigen die Gesten am Anfang mit mehr Kraft auszuüben als nötig. Trotzdem sollten hier Alternativen untersucht werden.

Die Probanden wurden am Ende des Tests gebeten auszuwählen, welches ihre bevorzugte Zeigemethode war. Dreimal landete die Gaze-Methode auf dem ersten Platz, die Indirekte-Methode zweimal. Nur ein Teilnehmer präferierte den Direkten-Modus, welcher bei vier Teilnehmern auf dem letzten Platz landete.

5.3 Schwächen der Anwendung

Alle Teilnehmer bemerkten eine Diskrepanz zwischen der optischen Zeigerüberlagerung mit den Objekten und dem Moment, in dem der Zeiger erkennt, wann ein Objekt vor ihm liegt. Weil sich die zweidimensionale Videoebene mitsamt dem Zeiger hinter allen dreidimensionalen Objekten befindet, muss sich der Anwender selbst direkt vor dem Trackingtarget positionieren, damit eine Überlagerung optimal funktioniert. Sieht der Anwender mit einem leichten Winkel auf das Target, kann es so aussehen als ob Zeiger und Objekt überlappen, der Zeiger befindet sich in Wahrheit aber nicht direkt hinter dem Objekt. Um dies zu lösen wird der Raycast, der von dem Zeiger ausgeht, in Zukunft direkt an den Anwender zurück gefeuert.

6 Fazit und zukünftige Arbeit

In dieser Arbeit wurde die Entwicklung drei unterschiedlicher Zeigemethoden für AR Anwendungen vorgestellt. Eine Studie unter Verwendung des Fragebogens aus ISO 9241-9 und der Talk-Aloud Methode mit sechs Teilnehmern hat Tendenzen und Ideen für die Weiterentwicklung der Methoden erbracht.

Keine der Methoden bereiteten den Probanden in der Anwendung Schwierigkeiten, was für eine einfache Erlernbarkeit der Konzepte spricht. Die untersuchte Methode des direkten Zeigens konnte allerdings wegen den hohen Ermüdungserscheinungen, vor allem im Arm, für weitere Projekte ausgeschlossen werden. Aufgrund der Ergebnisse wird die Idee des indirekten Zeigens sowie die Gaze-Methode speziell im Projektumfeld für Lernanwendungen weiterverfolgt und optimiert. In weiteren Arbeiten sollen auch erweiterte Tests zur Genauigkeit und Geschwindigkeit der Methoden durchgeführt werden.

7 Anerkennung

Das Projekt wird in der Förderlinie Digitale Medien in der beruflichen Bildung (DIMEBB) durch das Bundesministerium für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01PD14014A gefördert.

8 Literaturverzeichnis

[1] G. Bruder, F. Steinicke and K. Rothaus. Enhancing presence in head-mounted display environments by visual body feedback using headmounted cameras. In Proceedings of the 2009 International Conference on CyberWorlds (CW'09), S. 43–50. 2009.

[2] F. Haque, M. Nancel and D. Vogel. Myopoint: Pointing and Clicking Using Forearm Mounted Electromyography and Inertial Motion Sensors. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15), S. 3653-3656. 2015.

[3] N. Cournia, J. D. Smith, and A. T. Duchowski. Gaze- vs. hand-based pointing in virtual environments. In CHI '03 Extended Abstracts on Human Factors in Computing Systems (CHI EA '03), S. 772-773. 2003.

[4] S. A. Douglas, A. E. Kirkpatrick, and I. S. MacKenzie. Testing pointing device performance and user assessment with the ISO 9241, Part 9 standard. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '99), S. 215-222. 1999.

[5] S. Yoo, C. Parker. Controller-less Interaction Methods for Google Cardboard. In Proceedings of the 3rd ACM Symposium on Spatial User Interaction (SUI '15), S. 127-127. 2015.

[6] M. C. Cabral, C. H. Morimoto, and M. K. Zuffo. On the usability of gesture interfaces in virtual reality environments. In Proceedings of the 2005 Latin American conference on Human-computer interaction (CLIHC '05), S. 100-108. 2005.

[7] A. Monk, P. Wright, J. Haber and L. Davenport. Improving your human-computer interface: A practical technique. Prentice-Hall. London. 1993.

[8] N. Bevan. Usability is quality of use. Proceedings of the 6th International Conference on Human Computer Interaction. S. 349-349. 1995.

Cloud und die vertragliche Basis - die Zukunft von Service Level Agreements *

Aktuelle Ansätze zum Umgang mit Cloud-SLAs und Abbildung auf OpenStack

Dominik Waas
Reutlingen University
Dominik_Stefan.Waas@Student.Reutlingen-University.DE

Abstract

Cloud-Computing hat sich seit einigen Jahren in der IT-Branche etabliert, sodass einfache Anwendungen oder neue Entwicklungen meist auf Cloud-Technologie basieren. Das zunehmende Interesse von Unternehmen auch kritische Enterprise-Software als Cloud-Service anzubieten, führt zu neuen Herausforderungen. Bekannte Public Cloud-Angebote bieten nur eingeschränkt Zusicherungen und entsprechen nicht den Anforderungen für kritische Dienste. Unternehmen möchten daher eine Private Cloud mit individuellen Vereinbarungen (Service Level Agreements) und einem garantierten Leistungsstandard nutzen. Durch die dynamische Infrastruktur der Cloud und der Nutzung physikalischer Ressourcen durch verschiedene Kunden werden neue Konzepte zur Überwachung und Einhaltung von SLAs notwendig. OpenStack ist eine Open-Source Softwareplattform zur Bereitstellung von Cloud Ressourcen. Die Entwicklung und den Einsatz durch namenhafte IT-Unternehmen führt zu steigendem Interesse anderer Firmen OpenStack zur Realisierung von Private Clouds zu nutzen.
Die Arbeit erläutert die Grundlagen von Service Level Agreements und deren Inhalte und stellt die konkrete Problematik und Zielsetzung im Kontext ”Cloud” dar. Es werden aktuelle Ansätze zum Umgang mit Cloud-SLAs allgemein und im Rahmen von OpenStack betrachtet und ausgewertet. Die Erkenntnisse werden zusammengefasst und in einem Konzept zur Umsetzung einer SLA-Management-Erweiterung für OpenStack angewendet.

Schlüsselwörter

Cloud Computing, Service Level Agreement (SLA), Infrastructure-as-a-Service (IaaS), SLA Monitoring, SLA Enforcement, dynamic SLAs, Private Cloud, OpenStack, Ceilometer

CR-Kategorien

C.2.4 [**COMPUTER-COMMUNICATION NETWORKS**]: Distributed Systems - Distributed applications

1 Einleitung

Das Konzept ”Cloud” ist auch einige Jahre nach dessen Einführung ein aktuelles Thema der IT- und Kommunikationsbranche und erhält stets neue Entwicklungen und Nutzer. Besonders private Anwender nutzen vermehrt ”Software-as-a-

*

Betreuer Hochschule:
Prof. Dr.-Ing. Marcus Schöller
Marcus.Schoeller@Reutlingen-University.DE

Wissenschaftliche Vertiefungskonferenz
18. November 2015, Hochschule Reutlingen

Service" (SaaS) Angebote. Neue Softwareprodukte werden meist bereits als cloudfähige Anwendung entwickelt. Zunehmend steigt das Interesse von Unternehmen sog. "Enterprise-Software" von der traditionellen Client-Server-Architektur in eine verteilte Cloud-Anwendung zu transformieren. Auch wenn dies eine komplexe Aufgabe ist, wird sich eine höhere Agilität und Skalierbarkeit der Anwendung mit geringeren Ausfallzeiten (Downtime) und besseren Verwaltungsmöglichkeiten (Management) [17] erhofft. Bei dieser Entscheidung gibt es jedoch einige Bedenken. Da Ressourcen einer Cloud-Infrastruktur von mehreren Teilnehmern gleichzeitig genutzt werden (Multi-Tenancy), ergeben sich kritische Punkte in Bezug auf Sicherheit, Datenschutz und Ressourcenzusicherung. Allgemeine Vertrauensprobleme der Kunden oder Interessenten von Cloud-Angeboten werden weiterhin als größte Bedenken angeführt. [11] Zudem kann es eine schwierige Aufgabe sein vorhandene Anwendungen von klassischer, dedizierter Hardware in die Cloud zu migrieren. Dedizierte Systeme wurden meist mit dem Ressourcenbedarf für eine Maximalauslastung bzw. den Worst-Case bereitgestellt, sodass es dabei nie zu einem Engpass gekommen sein sollte. Eine Überspezifizierung des Ressourcenbedarfs von Cloud Services ist jedoch kontraproduktiv, da der Anbieter (Service Provider) Kapazitäten bereithalten muss, welche in den meisten Fällen nicht ausgelastet werden und dadurch mehr Kosten für den Kunden entstehen. Der Einsatz von Cloud-Technologie macht somit ein Umdenken erforderlich, sodass dem Anbieter mehr Freiraum bei der Ressourcenplanung zugestanden wird, sofern dieser die gewünschten Leistungen garantieren kann. Dies hat zur Folge, dass Cloud-Systeme mit weniger Kapazitäten bei geringeren Kosten und niedrigerem Energiebedarf betrieben werden können. [12]

Als vertragliche Basis zwischen einem Kunden und dem Service Provider dient ein sog. "Service Level Agreement" (SLA). In diesem werden unter anderem der Ressourcenbedarf, einzuhaltende Parameter und Verpflichtungen beider Parteien festgehalten. Die verfügbaren Public Cloud Angebote wie bspw. "Amazon AWS" oder "HP Helion Public Cloud" verfügen meist über ein vorgefertigtes SLA. Bei einer Bewertung nach [13] sind diese mangelhaft und effektiv für den Kunden nutzlos. Besonders Zusicherungen über die Verfügbarkeit (Availability) des Angebots werden stark eingeschränkt. Verfügbarkeiten werden oft nur für einzelne Regionen, nicht jedoch für ein Rechenzentrum oder Instanzen garantiert. Auch reklamierte Ausfallzeiten werden von den Anbietern nur geringfügig entschädigt, was in keiner Relation zu den tatsächlich entstandenem Schaden steht.

Der Bedarf an Sicherheit und individuellen SLAs mit Performance-Zusicherungen sind die Hauptgründe, weshalb besonders Enterprise-Kunden mit kritischen Anwendungen ein Private Cloud Angebot bevorzugen. [17] Das vereinbarte SLA mit dem Service Provider enthält detaillierte Parameter wie Verfügbarkeit oder Antwortzeiten. Bereits jetzt wird jedoch der Fokus nicht mehr auf einzelne Werte gelegt sondern auf die gesamte Produktqualität bzw. "Quality of Experience" (QoE) um letztlich dem Endkunden eine optimales Service Level zu bieten. [19]

OpenStack ist eine Open-Source Architektur bestehend aus Softwarekomponenten, die einen Zugriff und die Bereitstellung von Infrastruktur (Hardware) virtualisiert und skalierbar in Form von "Infrastructure-as-a-Service" (IaaS) ermöglichen. Das Projekt wird derzeit von großen IT-Firmen in der Entwicklung vorangetrieben, da sich so der vermehrte Einsatz von Private Cloud Architekturen erhofft wird, was wiederum einen gesteigerten Bedarf an Hardware und Serviceleistungen mit sich bringen würde. Der Trend zum Einsatz von OpenStack wurde durch namenhafte Firmen wie eBay, PayPal, Walmart oder BMW gesetzt, sodass mittlerweile über 740 Firmen die Komponenten im Einsatz haben. [7] OpenStack verspricht eine erweiterbare, flexible und elastische Cloud-Infrastruktur zu ermöglichen. Die Anwender erhoffen sich durch die offene Architektur einen Investitionsschutz und wollen einen sog. "Vendor Lock-in" vermeiden, da eine auf OpenStack basierende Infrastruktur von vielen verschiedenen Anbietern einheitlich bereitgestellt werden kann. [17]

Die im Folgenden dargestellte Arbeit beschäftigt sich mit dem Umgang von Service Level Agreements (SLAs) im Kontext von Cloud-Services. Die aus der Recherche gewonnenen Erkenntnisse werden auf die OpenStack Architektur angewendet und ein Konzept zur Überwachung von SLA-Parametern in OpenStack dargestellt. Im ersten Abschnitt werden die grundlegenden Elemente eines klassischen SLAs und der entsprechende Umgang damit betrachtet. Es wird auf die Besonderheiten in Bezug auf Cloud-Technologie und die daraus entstehende Problemstellung bzw. Zielsetzung eingegangen. Der darauf folgende Teil zeigt vorhandene Ansätze und Arbeiten im Bereich Standardisierung und Best Practices zu Cloud-SLAs. Es werden Arbeiten des European Telecommunications Standard Institute (ETSI), des TeleManagement Forums (TMForum), sowie des Open Grid Forums (OGF) einbezogen. Neben der Vorstellung des jeweiligen Ansatzes werden die Empfehlungen der einzelnen Arbeitsgruppen extrahiert. Der dritte Abschnitt geht konkret auf die OpenStack Architektur ein. Neben dem grundlegenden Aufbau des Systems wird auf thematisierte Ideen der OpenStack-Community und zwei Entwürfe zum Umgang mit SLAs innerhalb OpenStack eingegangen. Darauf folgend wird ein Konzept zur kontinuierlichen Überwachung eines Service Level Agreements (SLAs) vorgestellt. Abschließend werden die Erkenntnisse zusammengefasst und ein Ausblick zur weiteren Arbeit in diesem Bereich gegeben.

2 Grundlagen & Problemstellung

Der folgende Abschnitt erläutert die Grundlagen über SLAs und stellt die Bedeutung von SLA Management und die Besonderheiten im Cloud-Bereich dar. Es wird die konkrete Problemstellung aufgezeigt und die Zielsetzung verschiedener Parteien gesammelt.

2.1 Service Level Agreement (SLA)

Ein SLA wird als eine "Vereinbarung zwischen einem IT Service Provider und einem Kunden" (nach ITILv3 [8]) beschrieben. Es ist damit ein rechtlicher Vertrag, in welchem der Service Provider ein Service Level Versprechen macht. Es wird ein gemeinsames Verständnis aller Aspekte des Produkts, einzelner Rollen und entsprechende Verantwortlichkeiten ausgedrückt. Außerdem sind darin Attribute wie Verfügbarkeit, Wartbarkeit, Betrieb, Abrechnung und Strafen für nicht gelieferte Leistungen dokumentiert [19] [20]. Allgemein soll ein SLA die "Quality of Service" für einen Anwender des IT Systems definieren [12]. Es wird zwischen standardisierten "Off-the-Shelf" SLAs und individuellen Verträgen, welche meist bei kritischen Daten oder speziellen Anforderungen zum Tragen kommen, unterschieden. Neben der Definition von jedem gewünschten Service müssen zugehörige Metriken definiert werden, welche durch passende Auditierungs-Mechanismen und eine Überwachung der Werte überprüft werden können. Neben messbaren Werten gehören auch Elemente wie bspw. der Standort zur Datenablage um gesetzlichen Anforderungen zu entsprechen oder die konkrete Definition was unter "Verfügbarkeit" verstanden wird in ein SLA. [2] Eine detaillierte Auflistung einzelner Metriken mit Grenzwerten werden zur späteren Überwachung der Vereinbarung verwendet.

2.1.1 Service Level Objective (SLO)

Unter einem Service Level Objecitve (SLO) wird meist ein objektiv messbarer Zustand oder eine Charakteristik (bspw. Datendurchsatz oder Ausführungsdauer) verstanden. SLOs können auch aus einzelnen Metriken

zusammengesetzt sein. Es wird empfohlen SLOs auch nach ihrer Relevanz zu bewerten [2], da zum Beispiel die Verfügbarkeit eines System wichtiger als Antwortzeiten einzelner Dienste sein kann. In der Arbeit des TMForums [20] werden SLOs auch als "Service Level Goals", also freiwillige Zielvorgaben ohne verbundene Strafen bei Nichterfüllung, beschrieben. Zudem können sog. "Business Level Objectives" (BLOs) definiert werden, welche nicht direkt den Service spezifizieren, jedoch bei der Auswahl benötigter Metriken berücksichtigt werden sollten.

2.1.2 SLA Management

SLA Management deckt den gesamten "Customer Experience Lifecycle" ab, also den gesamten Zeitraum von der Inbetriebnahme über die Nutzung eines Dienstes (inklusive Kundenservice und Abrechnung) bis hin zur Beendigung des Vertrages. Bestandteile sind unter anderem die Vertragsdefinition, Verhandlung und Überwachung der Einhaltung definierter Richtlinien und die Verwaltung der Quality-of-Service Parameter. SLA Management lässt sich damit in zwei wesentliche Phasen aufteilen [20]. Ein Teil betrifft die Verhandlung des SLAs selbst. Der andere Teil beschäftigt sich mit der, am besten in Echtzeit erfolgenden, Überwachung (Monitoring), dass definierte Parameter eingehalten werden.

2.1.3 SLAs im Kontext Cloud

Bei der Entwicklung eines SLAs in Bezug auf die Nutzung von Cloud-Resourcen ist es besonders schwierig das passende Service Level zu finden, um ein zuverlässiges System zu erhalten. Kunde und Service Provider sollten sich auf ein Service Level einigen und im Laufe des Betriebs die Bereitstellung anhand von Messergebnissen überprüfen. Neben der üblichen Definition von Metriken und Grenzwerten, sollte besonders der Eigentümer und die Lesbarkeit des Datenstandes sichergestellt werden und ein detaillierter Überblick über die Cloud-Infrastruktur, sowie die vom Service Provider einzuhaltenden Sicherheitsstandard gegeben sein. Ebenfalls sollte sich der Kunde ein Recht auf Auditierung vorbehalten, um eine Einhaltung (Compliance) des Vertrags prüfen zu können. Da es sich bei Cloud-Angeboten um stark dynamische Umgebungen handelt, macht eine ständige Neubewertung der vereinbarten SLAs und ggf. eine Änderung der Vereinbarungen Sinn. [1]
Das Service Level Management beschäftigt sich in Cloud-Umgebungen zunehmend damit, wie Performance-Informationen gesammelt und genutzt werden können. Cloud Service Provider können mit den passenden Metriken technische Entscheidungen, bspw. zur Infrastruktur, treffen. Kunden (Cloud Consumer) hingegen können diese Informationen dazu nutzen um die angebotenen Dienste eines Providers zu evaluieren. [2]

2.2 Problemstellung

Public Cloud Angebote wie Amazon AWS oder Microsoft Azure bieten meist nur ein vorgefertigtes SLA mit verschleierten Verfügbarkeits-Zusicherungen. So werden oft nur die Verfügbarkeit einzelner Schnittstellen garantiert, jedoch nicht ein Ausfall welcher bspw. durch eine fehlende Konnektivität zum Internet zustande kommt. Neben den geringfügigen Erstattungen im Fehlerfall sind solche Verfügbarkeitsgarantien für einen Endbenutzer nicht tauglich. Die konkreten Bedingungen zur Beendigung eines Vertrags treten auch immer weiter in den Hintergrund, da die meisten Angebote ohnehin nur noch auf monatlicher Basis bzw. nutzungsabhängig "Pay-as-you-go" abgerechnet werden.
Das Versprechen eines Anbieters bzgl. Performancezusicherungen reicht nicht aus - der Kunde muss überzeugt werden, dass der Service Provider die gewünschte Leistung auch liefern kann. Cloud-Kunden werden virtualisierte Ressourcen anstelle physikalischer Hardware (bare metal) zur Verfügung gestellt. Dazu wird unter anderem eine Adaption bekannter Metriken für Cloud Computing notwendig, da Ressourcen dynamisch zugeordnet werden können und eine kontinuierliche Veränderung der Kapazitäten stattfindet. [11] Ressourcenintensive Anwendungen anderer Kunden auf derselben Hardware (sog. "Noisy Neighbors") können außerdem zu Performanceeinbußen führen. Enterprise-Kunden suchen jedoch eine Möglichkeit kritische Anwendungen in Private Clouds unter bewiesenen bzw. stark kontrollierten Leistungsgarantien zu betreiben. [17]
Die vorhandenen Basis-SLAs sind für kritische Anwendungen nicht ausreichend, da ein zuverlässiges Monitoring mit Erkennung und Reaktion von nicht erfüllten Leistungsangaben dazu notwendig ist. Die meisten Arbeiten betrachten derzeit eher nichtfunktionelle Eigenschaften wie die Sicherstellung von Anti-Affinity-Regeln oder Redundanzen. [12] Eine Überspezifizierung der Parameter ist für dieses Problem keine Lösung, da Ressourcen meist unnötig bereitgehalten werden und unnötig Kosten entstehen. Man kann dabei von einer "Dedicated Cloud" sprechen, was jedoch den Sinn und vermutlich die Business Objectives des Kunden verfehlen.
Als eine der komplexen Aufgaben zeichnet sich das Finden von passenden "Key Quality Indikatoren" (KQIs), welche im folgenden Abschnitt detaillierter erläutert werden. [19] Es wird klar, dass Cloud Computing ohne passendes Service Management, Lifecycle-Management, Metering, Monitoring und sauber definierte SLAs nicht möglich ist. [2] Die Empfehlung vorhandene Standards zu erfüllen erweist sich als schwierig, da es bisher keine vollständigen und final veröffentlichten Leitfäden zum Umgang mit der Thematik gibt. Eine Hauptaufgabe zur Lösung des Problems ist es zuverlässig Echzeitdaten zur Überprüfung der SLA-Metriken zu sammeln und diese methodisch und basierend auf einem Standard zu bewerten und entsprechend einer Strategie (Policy) automatisiert zu verarbeiten. [18]

2.3 Zielsetzung

Die Cloud-Architektur und Anforderungen der Kunden erfordern einen transparenten und zuverlässigen Umgang der Cloud Provider mit SLAs. [2] Das Vertrauen des Nutzers kann nicht durch einfache Versprechen gewonnen werden, sondern muss durch Überzeugung mit der bereitgestellten Infrastruktur und den eingesetzten Methoden zur Sicherstellung der vertraglichen Vereinbarung hergestellt werden. Der Cloud Provider muss eine bessere Kenntnis über die Auslastung der Systeme durch den Kunden haben, um eine bessere Kapazitätsplanung und Lieferung der geforderten Service Levels durchführen zu können [12]. Der Kunde erwartet zudem eine sog. "Service Awareness" mit proaktivem Handeln, d. h. der Anbieter ist für eine kontinuierliche Überwachung und vorzeitigen Eingriff bei kritischen Entwicklungen zuständig. [9]
Die Anforderung einer maschinenlesbaren SLA-Definition wird ausnahmslos gestellt, um eine automatisierte und dynamische Verhandlung und Abwicklung verschiedener Prozesse zu ermöglichen. Eine maschinenlesbares SLA kann bspw. vom Kunden genutzt werden um den günstigsten oder sichersten Anbieter für ein Angebot zu finden [2] oder vom Provider um eine automatische Ressourcenplanung vorzunehmen. [12] Das Monitoring sollte dynamisch und kontinuierlich Arbeiten und vertrauenswürdig präzise Daten in Echtzeit liefern. Dabei sollten sowohl quantitative als auch qualitative Metriken, wie die Wahrscheinlichkeit oder Anfälligkeit eines Ausfalls, mit einbezogen werden. [18] Dem Kunden sollte eine Auditierung durch einen unabhängigen Drittanbieter ermöglicht werden [11], indem klar definierte Methoden zur Erfassung von Metriken und passende Ausgaben zur Berichterstattung genutzt werden. Im optimalen Fall kann der Kunde die "Quality of Service" (QoS) und wichtige "Key Quality Indikatoren" (KQIs) der einzelnen Services bzw. Service Provider in einer strukturierten Übersicht (Dashboard) einsehen, wie diese über den übergreifenden Rahmenvertrag der Dienstleistung (Master Service Agreement ("MSA")) bzw. das spezifische Service Level Agreement festgelegt wurden. [20]
Auch wenn der Fokus der jeweiligen Ar-

beitsgruppe ein anderer ist, so wird besonders die Forderung von einem Standard und die maschinenlesbare SLA-Definition deutlich. Ein automatisierter und zuverlässiger Prozess zur Überwachung von SLAs wird erst mit diesen Bestandteilen ermöglicht.

3 Aktuelle Ansätze & Arbeiten

Derzeit gibt es verschiedene Arbeitsgruppen, welche sich mit der Standardisierung von Cloud-SLAs und den zugehörigen Prozessen wie Management, Monitoring und Enforcement (Sicherstellung der Einhaltung) beschäftigen. Das Cloud Standard Customer Council (Cloud Council) beschäftigt sich derzeit mit der Ausarbeitung eines SLA Leitfadens. [1] Weitere Arbeiten kommen vom NSF Center for Cloud and Autonomic Computing oder von den EU-Arbeitsgruppen OPTIMIS oder SLA@SOI. Projekte die sich besonders der Zuordnung zwischen abstrahierten SLA-Metriken (Quality Parameter) auf einfache Ressourcen-Metriken beschäftigen sind unter anderem EVEREST und DeSVI. [12] Ein weiteres Bestreben der Cloud Security Alliance (CSA), ENISA, Cloud Audit (A6) oder OCCI ist es einen Standard für Cloud Sicherheit und der Risikobewertung von Angeboten zu erstellen. Auch wenn die Arbeiten sich teilweise mit unterschiedlichen Spezialisierungen beschäftigen, so geben alle die Empfehlung für eine Automatisierung von SLAs mit Hilfe einer SLA-spezifischen Sprache bzw. in maschinenlesbarer Form ab. [11]
Im Folgenden werden konkret die Ansätze und Arbeiten des ETSI, TMForum Enterprise Cloud Leadership Council (ECLC) und das OGF WS-Agreement betrachtet.

3.1 ETSI NFV Spezifikation

Die ETSI NFV Spezifikation [10] ist im speziellen Kontext der Virtualisierung von Netzwerkfunktionalitäten für die Telekommunikationsindustrie entstanden. Dennoch können daraus grundlegende Vorgehensweise zur Verwendung von Metriken entnommen werden. Im Folgenden werden die einzelnen Elemente einer Messung definiert und exemplarisch die Erarbeitung einzelner Metriken aufgezeigt.
Unter einer Messung (Measurement) versteht man eine Reihe an Aktionen, um einen Messwert zu bestimmen bzw. ein Ergebnis zu erhalten. Ein Messpunkt (Measurement Point) ist ein physikalischer oder logischer Punkt der Beobachtung. Unter einer Metrik (Metric) wird die exakte Bedeutung eines Messwerts, also die Standard-Definiton festgelegt. Eine abgeleitete Metrik (Derived Metric) nutzt Werte anderer Metriken, um eine Angabe zu machen. Bspw. kann die einzelne Metrik ”Paketverzögerung” in die Ermittlung der ”Paketübertragungs-Performance” einbezogen werden. Unter einem Parameter versteht man einen Eingabefaktor in Form einer Variable für die Definition einer bestimmten Metrik. [10]
Die Spezifikation teilt die definierten Metriken in drei Service Metrik Kategorien auf: Geschwindigkeit (Speed), Fehlerfreiheit (Accuracy) und Zuverlässigkeit (Reliability). Zudem lassen sich die Metriken innerhalb einzelner Bereiche (Steps) im Lifecycle einordnen, wie zum Beispiel Orchestration, Aufbau oder Betrieb. Zur Einordnung wird oft auf die effektiv nutzbare Betriebszeit, das ”Useful Life Timeframe”, verwiesen. Durch diese Angabe ist eine klare Identifizierung der Zuständigkeit möglich. Metriken, wie eine Provisioning-Latenz oder Instanzen, die als ”Dead-on-Arrival” (DOA) markiert werden, sind über das Orchestrations Service Quality Management abzudecken. Blockierte (Stalls) oder frühzeitig beendete (Premature Release) VMs treten nach Beginn des Useful Life Timeframes ein und sind damit dem Betrieb zuzuordnen. [10]
Das Dokument beschreibt verschiedene Messwerte in detaillierter Tiefe. Exemplarisch kann hierzu die Definition eines ”VM Stalls” angeführt werden. Dies bedeutet, dass keinerlei Ausführung in einer Instanz während dieses Zeitfensters möglich ist. Neue Anfragen und wartende Aufträge werden nicht bearbeitet. Eine Messung sollte den Zeitraum zwischen dem Stoppen der VM und dem Fortfahren (bspw. durch eine Live Migration) erfassen. Brauchbare Metriken könnten bspw. die Dauer dieses Vorgangs oder die Häufigkeit sein. An die Messung sollte eine Aktion gekoppelt werden. Wenn die Dauer eines VM Stalls einen vorgegebenen Maximalwert (Threshold) überschreitet, soll ein Hochverfügbarkeitsmechanismus aktiviert werden. Es wird auf eine neue VM Instanz gewechselt und der Vorgang wird als vorzeitige Beendigung der VM gewertet. Beim Aufbau von Metriken sollte außerdem der konkrete Einfluss einer einzelnen Messung auf den Gesamtwert klar werden. Die Rate von VMs, welche als ”Dead-on-Arrival” (DOA) gekennzeichnet werden, beeinflusst bspw. die Metriken wie elastisch ein Service ist (Scale-Out) und wie gut die Reparaturfähigkeit ist. Ein anderes Beispiel wäre Jitter bzw. die Variation der Paketverzögerung. Dies führt zu einer Instabilität der Kommunikation, was sich besonders im Telekommunikationsbereich auf die gesamte Service Latenz und letztlich auf die Quality-of-Service (QoS) des Endkunden auswirkt. [10]
Die ETSI NFV Spezifikation geht besonders detailliert auf die strukturierte Auflistung einzelner Messungen und die Zusammensetzung zu Metriken ein. Es wird empfohlen nicht pauschal alle möglichen Werte zu messen, sondern Messungen anzustreben, welche besonders Einfluss auf die Servicequalität des Kunden haben. Dazu kann ein Charakterisierungs-Plan erstellt werden mit Angaben welche Werte man wo und wie oft messen sollte. Um die Messaktivität steuern und überwachen zu können sollte das Management System eine Möglichkeit zur Kontrolle und Berichterstattung bieten. Es empfiehlt sich auch eine Auditierung durch Abgleich der gemessenen Werte zwischen Provider und Kunde zu ermöglichen um so ggf. Inkonsistenzen feststellen zu können. Im SLA sollte zudem definiert werden wie die Messergebnisse zu vergleichbaren Berichten und Statistiken zusammengefasst werden. Eine Aggregation von detaillierten Einzelmesswerten über einen längeren Zeitraum kann sinnvoll sein, um bspw. Trendanalysen der Servicequalität durchführen zu können. [10]
Zusammenfassend lässt sich die dargestellte Spezifikation besonders zur strukturierten Auflistung der verwendeten Messwerte und daraus entstehenden Metriken nutzen. Eine Klassifizierung der Metriken kann außerdem dazu genutzt werden, um einen schnellen Einblick zu erhalten, welche Kategorie und welcher Lifecycle-Schritt am meisten von Abweichungen beeinträchtigt sind.

3.2 TMForum SLA Management

Das TeleManagement Forum beschäftigt sich ausführlich mit den Themen ”SLA Management” und konkret mit der Umsetzung von SLAs in einem komplexem Cloud-Ecosystem mit mehreren Teilnehmern. Das SLA Management Handbook [19] gibt eine grundlegende Definition von beteiligten Rollen, verwendeten Metriken und dem Aufbau des SLAs. Das Dokument zur End-2-End-Umsetzung des SLA Managements [20] stützt sich auf das vorhergehende Handbuch und stellt eine Verfahrensweise für die Cloud-Umgebung vor.
Für ein passendes Verständnis ist die Definition von Organisation, Akteur (Actor) und den jeweiligen Rollen im SLA Management notwendig. Eine Organisation ist eine rechtliche Entität wie eine Firma, Person oder auch eine Gruppe von Personen. Ein Akteur kann als Teil einer Organisation mit bestimmten Zuständigkeiten für einen Bereich gesehen werden. Dieser kann ein oder mehrere Rollen haben. Eine Rolle ist ein Satz an Aktivitäten, bspw. Service Provider, Kunde, Benutzer oder Integrator. [19] Dementsprechend stehen zwei Organisationen in einer Beziehung zueinander, welche sich über ein SLA klar definieren lässt.

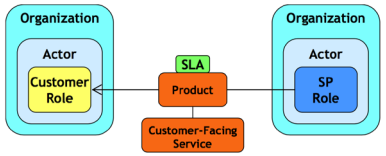

Abbildung 1: Beziehung zwischen Service Provider und Kunde nach TMForum ([19][S. 12])

Das SLA Management Handbook nimmt eine starke Strukturierung der Vereinbarung zwischen Kunde und Provider vor. Als Grundlage dient das allgemeine Business Agreement (BA). Das darin enthaltene SLA ist stets an ein Produkt gekoppelt und definiert welche Messmethoden und Berichtsprozesse genutzt werden sollen.

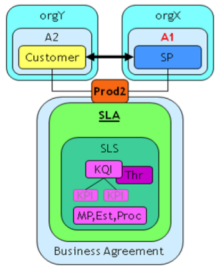

Abbildung 2: Produkt mit zugehörigem SLA und einzelnen Inhaltselementen ([19][S. 23])

Um eine Abtrennung der Definition von Performance-Indikatoren und Zielwerten zu erhalten wird die ”Service Level Specification” (SLS) eingeführt. Diese soll definieren welche Parameter wie, wo und wann gemessen werden (Applicability) und mit welchen Werten (Thresholds) die Ergebnisse aufgewogen werden. Die SLS besteht konkret aus einzelnen Key Quality Indikatoren (KQIs) bzw. Key Performance Indikatoren (KPIs) und Parametern wie Grenzwerte (Thresholds). Es können mehrere Grenzwerte in einem Dokument definiert werden um verschiedene Stufen des Service Levels auszudrücken (bspw. Verfügbarkeitsgarantie auf Bronze-, Silber- und Gold-Level). [19]

Der Provider beschäftigt sich vorwiegend mit den Key Performance Indikatoren (KPIs). Ein KPI ist eine konkrete Metrik welche die Performance einer oder mehrerer Ressourcen wiedergibt. Ein Kunde ist jedoch primär an den Key Quality Indikatoren (KQIs) interessiert. Diese Metriken können aus anderen KQIs, verschiedenen KPIs oder direkten Messungen bestehen und geben Auskunft über die Performance eines Produkts. [19] KQIs können einfach bspw. mit Grenzwerten gestaltet sein, jedoch auch aus mathematischen Formeln wie Korrelation oder einer Summe der Mediane [20] aufgebaut sein.

Für eine vollwertige Cloud-Anwendung werden heutzutage verschiedene Services von unterschiedlichen Anbietern in einem Cloud-Ecosystem kombiniert. Oft ist dabei keine direkte Unterscheidung zwischen Kunde und Anbieter eines Services möglich, da ein Produkt durch die Kombination mehrerer Services entstehen kann. Ein Endkunde fordert jedoch einen ”Single Point of Accountability” [20], also einen Anbieter mit einem SLA, welchen er im Fehlerfall zur Verantwortung ziehen kann. Die dadurch entstehende Komplexität benötigt ein sauberes SLA End-2-End Management, welches bspw. durch einen ”Lead Service Provider” übernommen werden kann. Für Aufgaben wie die Rechnungsstellung oder eine Auditierung der KQIs oder die Sicherstellung der SLA Parameter kann auch ein unabhängiger Drittanbieter eingesetzt werden.

Das TMForum TR178 Dokument erweitert das vorhergehende SLA Management Handbook um die Betrachtung des Cloud-Szenarios. Es wird vorgeschlagen SLA Metriken in zwei Kategorien aufzuteilen. Business Metriken dienen der nutzungsbasierten Abrechnung (Billing), technische Metriken hingegen sollen für das Monitoring eingesetzt werden, um die Einhaltung des SLAs zu prüfen. Wie auch in der ETSI NFV Spezifikation müssen Metriken definiert werden und bspw. in Abhängigkeit vom Modell (IaaS, PaaS oder SaaS) oder Typ des Services (Computing, Netzwerk, Speicher) kategorisiert werden. Metriken sollten mit entsprechenden Werten wie Minimum, Maximum, Standardwert gekoppelt werden und entsprechende Konsequenzen bei Abweichungen der Werte vermerkt sein. [20]

Die Arbeit schlägt vor, dass jeder Service Provider einen Service Katalog für Kunden und Entwickler anbietet, in welchem angebotene Dienste auf Basis eines standardisierten ”Serivce Templates” beschrieben werden. Dabei soll versucht werden vorhandene Industriestandards zu adaptieren. Eine feste Vorlage kann die verschiedenen Interessen der Parteien abdecken. Der Provider kann ein Service Template nutzen, um intern die Überwachung und Sicherstellung der SLA-Metriken zu automatisieren. Der Kunde kann Business und Service Objectives verschiedener Anbieter vergleichen und kontinuierlich das Service Level des Dienstes evaluieren. [20]

Der End-2-End Cloud SLA Management Leitfaden empfiehlt vorhandene Standards und Best Practices von Kunden und Cloud Providern zu betrachten und in einen standardisierten Prozess zu integrieren. Es muss eine Vorlage geschaffen werden wie ein Master Service Agreement, Cloud SLAs, BLOs, SLOs und die konkreten Metriken definiert werden. Für einen einheitlichen und automatisierten Austausch dieser Informationen soll ein Servicekatalog mit Service Templates und Managementberichten genutzt werden. Für die Verwendung von Service Templates sollte ein Taxonomie-Standard mit verpflichtenden Inhalten wie bestimmte Metriken und Parameter geschaffen werden, um eine Vergleichbarkeit zu gewährleisten. Je nach Industrie können die Inhalte erweitert werden. Die Definition des SLAs, die damit verbundenen Richtlinien und der Verhandlungsprozess müssen möglichst flexibel gestaltet sein [20], um die dynamische Cloud-Umgebung zu unterstützen.

Der Ansatz der TMForum Arbeitgruppe zeigt klar die Forderung nach konsistenten Methoden zum Umgang mit Cloud-SLAs. Als Lösung werden hier standardisierte Vorlagen, welche auf vorhandenen Industriestandards basieren, vorgeschlagen.

3.3 OGF Web-Service-Agreement

Die Web Services Agreement Specification (WS-Agreement) [15] beschreibt ein vielseitig einsetzbares Protokoll um eine Vereinbarung zwischen zwei Parteien zu finden. Die Nutzung zur Abbildung maschinenlesbarer SLAs wird von mehreren Arbeitsgruppen [20] [12] und dem Open Grid Forum selbst vorgeschlagen. Die Spezifikation enthält im Wesentlichen Schemata zur Spezifikation von Vereinbarungen und entsprechenden Vorlagen, sowie Informationen wie die Vereinbarungen technisch abgewickelt werden sollen. Mittels XML soll ein vorgefertiges Schema mit Feldern und Werten für die einzelnen SLA-Parameter genutzt werden, um Prozesse rund um das SLA Management zu automatisieren [20].

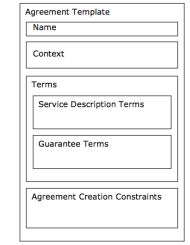

Abbildung 3: Aufbau - Agreement Templates ([15][S. 29])

3.3.1 OGF Web-Service-Agreement Negotiation

Die WS-Agreement Negotiation [16] erweitert die vorhergehende Spezifikation um den Prozess der SLA-Verhandlung. Der Standard nutzt ein ausgefülltes Service-Template (XML), um den Verhandlungsprozess zwi-

schen Kunde und Provider über einen Satz an Schnittstellen zu automatisieren.
Der Kunde nutzt das vom Provider zur Verfügung gestellte SLA Template und füllt dieses entsprechend seiner Serviceanforderungen aus. Das Dokument wird als sog. ”Offer” an den Service Provider gesendet, welcher die Anfrage auswertet und ein entsprechendes Angebot mit konkreten Werten (wie bspw. zugesicherten Metriken) zurücksendet. Der Kunde kann das Angebot nun Annehmen oder Ablehnen. Eine Neuverhandlung kann über neue Angebote stattfinden, solange bis der Kunde einem Angebot zustimmt. [20]
Die Akzeptanz der OGF WS Spezifikationen zeigt sich anhand steigender Zahl der Arbeiten und Implementierungen, welche auf dem Standard aufbauen. [12]

4 Anwendung auf OpenStack

Nachdem eine Betrachtung aktueller Ansätze und Arbeiten zum Umgang von Cloud-SLAs betrachtet wurde, wird im Folgenden auf die OpenStack Architektur und Arbeiten konkret in diesem Kontext eingegangen. Anschließend erfolgt die Konzeption einer möglichen Umsetzung eines Werkzeugs zur Überwachung von SLA-Metriken auf Basis vorhandener OpenStack Komponenten.

4.1 OpenStack Architektur

OpenStack ist eine Python-basierte Softwareplattform zur Bereitstellung von Cloud Computing Ressourcen. Das Open-Source Projekt wird von großen Firmen wie HP, NEC oder Red Hat vorangetrieben, wobei NEC derzeit am meisten Quellcodezeilen zum Projekt beigetragen hat. Die Programmierschnittstellen (APIs) von OpenStack sind mit Amazon EC2 und S3 kompatibel, sodass für Neuanwender einen reibungslosen Umstieg von Amazon Web Service auf OpenStack vornehmen können. [6]

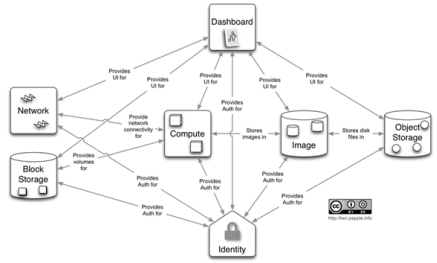

Abbildung 4: OpenStack Architektur (vereinfacht) ([3])

Die Architektur teilt sich in einzelne Komponenten auf, welche jeweils eine spezielle Aufgabe bzw. Dienstbereitstellung übernehmen. Alle Module haben einen Projektnamen zur einfachen Wiedererkennung. ”Keystone” wird für die Authentifizierung bei fast allen Vorgängen verwendet und ist damit sehr zentral in das System eingebunden. Bevor eine API-Anfrage bearbeitet wird, überprüft Keystone, ob der Benutzer über die entsprechenden Berechtigungen verfügt. Unter ”Nova” werden alle Compute-Komponenten zusammengefasst, um CPU- oder Arbeitsspeicher-Ressourcen zur Verfügung zu stellen. ”Glance” stellt für die Instanziierung von VMs eine Reihe an Images zur Verfügung. Persistenter Speicher kann über die Projekte ”Swift” in Form eines Objektspeichers oder ”Cinder” in Form von Block Storage bezogen werden. Einzelne VMs werden über die Netzwerkverwaltung ”Neutron” konnektiert. Eine skriptbasierte Orchestrierung der gewünschten Vorgänge kann mittels ”Heat” umgesetzt werden. [3] Die genannten Projekte bilden den Kern der OpenStack-Welt, wobei es neben den IaaS-Angeboten auch SaaS-Softwarekomponenten wie ”Trove” für Datenbanksysteme oder ”Designate” für die DNS-Bereitstellung gibt. Als unterstützende Services werden unter anderem die Verwaltungsoberfläche ”Horizon” und das Telemetrie-Modul ”Ceilometer” gewertet. [5]

Die Entwicklung von OpenStack findet in einzelnen Teams statt, welche weitestgehend selbstbestimmt arbeiten. Grundsätzlich kann jeder Entwickler sich an OpenStack beteiligen und seine Ideen und Konzepte über einen sog. ”Blueprint” einreichen. Die Konzeption und Implementierung wird von bestehenden Entwicklern beobachtet und in einem halbjährlichen ”Design Summit” der Vorschlag öffentlich diskutiert. [5][How to Contribute]
Für den Betrieb einer OpenStack Cloud wird eine Linux-Distribution wie openSUSE, Red Hat oder Ubuntu benötigt und eine Basis-Architektur aus Controller und Networking Node empfohlen. Je nach gewünschter Kapzität und Funktionsumfang des Clusters können beliebig viele Compute, Storage oder weitere Nodes hinzugefügt werden. [4][Architecture]
Für den Fokus der Arbeit sind besonders das Telemetrie-Modul ”Ceilometer” für ein SLA-Monitoring, sowie die Orchestrierungs-Komponente ”Heat” für ein SLA-Enforcement interessant.

4.2 Konzepte der OpenStack Community

Durch das zunehmend steigende Interesse von Unternehmen auch kritische Enterprise-Anwendungen mit OpenStack zu betreiben muss sichergestellt werden, dass die Service Level Anforderungen erfüllt werden können. Aktuelle Diskussionen zur möglichen Umsetzung von SLA-Monitoring und SLA-Enforcement beschäftigen sich besonders mit dem Scheduling einzelner Instanzen unter der Verwendung von Heat [8], Kapazitätsplanung [12] und sicherheitsrelevanten Überwachung von Ressourcen [17].
Um einen SLA-konformen Betrieb der Ressourcen zu erzielen kann bspw. in das Scheduling von Nova über neue Filter und Gewichtungsmethoden eingegriffen werden. Auch in Verbindung mit Heat kann die Orchestrierung bereits jetzt Richtlinien (Policys) nutzen um bspw. Vorgaben wie Anti-Affinity-Regeln über virtuelle Gruppierungen von Servern abzudecken. Allgemein werden jedoch mangelnde Flexibilität und Agilität als Defizite des Schedulers genannt. [8] Von Entwicklerseite wird häufig empfohlen, dass in die Prozesse von OpenStack anhand von SLA-Parametern eingegriffen werden sollte - es werden jedoch nicht weiter die vorhergehenden Schritte zur Überwachung dieser Parameter erläutert.
Im Folgenden werden zwei betrachtete Ansätze detailliert dargestellt.

4.2.1 Anwendung dynamischer SLAs

Die Arbeit von [12] stellt einen möglichen Umgang mit dynamischen Service Level Agreements in OpenStack vor. Ziel ist es vorhandene Enterprise-Anwendungen in eine OpenStack-Umgebung zu überführen und die genauen Leistungsanforderungen durch eine dynamische SLA-Architektur zu ermitteln. Der Kunde muss somit die genauen Metriken zur Sicherstellung seines Service Levels nicht kennen, sondern kann die Anwendung nach einer Lernphase durch Monitoring in eine SLA-kontrollierte Umgebung umziehen. Der Provider der OpenStack-Cloud soll dadurch die verschiedenen Ressourcenauslastungen der Kunden analysieren können und eine ausgeglichene Auslastung der gesamten Infrastruktur erreichen können.
Der Ansatz unterscheidet zwischen verschiedenen Belastungsarten. ”Best Effort” bezeichnet Lasten die keinem SLA unterliegen und je nach Ressourcenverfügbarkeit ausgeführt werden. ”Throttled” sind Lasten mit einem fest zugesicherten Minimum an Ressourcen. ”Load Migration” Lasten sind Prozesse welche ggf. auf andere Server migriert werden, wenn bestimmte Grenzwerte erreicht werden. Ob eine Migrierung vom Prozess unterstützt wird ist individuell, sodass ggf. auf den Lasttyp (Reserverkapazität) zurückgegriffen werden muss. Ein weiterer Lasttyp ist ”Preemption”, welcher meist unkritische Prozesse darstellt. Diese Prozesse können jederzeit ohne Vorwarnung von System beendet und die Ressourcen anderweitig verwendet werden.[12]

Das Konzept sieht zwei neue Module in der Architektur vor. Die zentrale Verwaltungskomponente ”Admission Control” soll die gesamte Cloudkapzität und die aktuell verfügbaren Ressourcen verwalten. Dem Modul sind sämtliche Kapzitäten, sowie die Cluster-Topologie mit allen Server und Netzwerkkomponenten bekannt. Ein Kunde kann bei dieser Komponente eine neue Instanz mit Performance-Parametern (bspw. aus einem SLA) beantragen. Dazu wird der WS-Agreement Negotiation Standard genutzt. Entsprechend der SLA Anforderungen wird ein passender Host ausgewählt und die Instanz durch Nova erstellt. Um einen akkuraten Überblick über die verfügbaren Ressourcen zu erhalten, agiert Nova kontinuierlich mit der Admission Control und einem ”SLA Manager”. [12]
Der SLA Manager soll vereinbarte SLA-Parameter über das Monitoring von Ceilometer erfassen und analysieren. Basierend auf einem Regelsatz werden die Messwerte evaluiert und mit den Sollwerten verglichen. Die Evaluation soll vorhergegangene Messwerte einbeziehen und Tendenzen für einen möglichen Verstoß des SLAs frühzeitig erkennen. Die Arbeit geht nicht weiter auf die tatsächliche Umsetzung des SLA-Enforcements ein. Nachdem ein Verstoß gegen SLA-Parameter erkannt wurde, werden verschiedene Szenarien zum Umgang damit vorgeschlagen. Es kann eine einfache Benachrichtigung ausgelöst werden und manuelles Handeln erforderlich sein oder weitere Schritte wie die Migration auf einen anderen Server, die Zuverfügungstellung von weiteren Ressourcen (Scale-Up) oder im letzten Fall die Neuverhandlung eines SLAs über den WS-Agreement Negotiation Prozess. [12] Die Arbeit wendet für das vorgesehen Monitoring unterschiedliche Technologien an, so wird bspw. auch vorgeschlagen auf dem Gast-Betriebssystem entsprechende Überwachungsprozesse zu nutzen. Dieser Vorschlag ist jedoch bei reinen IaaS-Angeboten nicht mit dem Kunden vereinbar.

4.2.2 Intel Service Assurance Technology

Das White Paper von Intel [17] stellt die sog. ”Service Assurance Technology” vor. Ziel der Erweiterung ist es den Cloud Service Katalog bzw. Flavors (vordefinierte VM-Konfigurationen) um Enterprise-SLA-Parameter zu erweitern, wie bspw. Performancegarantien, I/O Durchsatz oder die Cache-Nutzung. Durch kontinuierliches Monitoring sollen Verletzungen des SLAs erkannt und berichtet werden. Dem Kunde kann damit ein ”SLA Compliance Report” erstellt werden, was die Vertrauenswürdigkeit beim Kunden steigern soll. [17]
Vorhandene Dienste wie Ceilometer oder der Nova-Scheduler werden mit einem ”Service Assurance Controller” verbunden. Zusätzlich wird ein Agent auf den einzelnen Compute Nodes installiert um ”Deep Telemetry” wie bspw. Cache-Überwachung zu betreiben. Der Ansatz verfolgt die Empfehlung den Kunden durch methodisches Vorgehen vom Serviceangebot zu überzeugen. Compute Nodes können als ”vertrauenswürdig” zertifiziert werden, sobald sie eine Reihe an Überprüfungen (BIOS-Rootkits, Hypervisor-Typ, uvm.) bestanden haben. Zudem wird versucht agressive VMs und Ressourcen-Missbrauch (sog. ”Noisy Neighbors”) zu identifizieren und damit dem Kunden die Bedenken einer Ressourcenteilung mit anderen Teilnehmern (Multi-Tenant-Architektur) zu nehmen. [17]. Intel stellt mit der Technologie die fortgeschrittenste Arbeit in diesem Bereich dar. Die Nutzung vorhandener Werkzeuge von OpenStack und die Integration von SLA-Parameter in die jeweiligen Instanzkonfiguration stellen einen praktikablen Ansatz dar. Die Auswertung von Metriken und entsprechende Reaktionen im Fehlerfall werden von der proprietären Komponente, dem ”Service Assurance Administrator”, übernommen. Intel vertreibt dieses Produkt kommerziell und kostenpflichtig, sodass eine Nutzung der Erweiterung im Sinne des Open-Source-Gedankens von OpenStack nicht möglich ist.

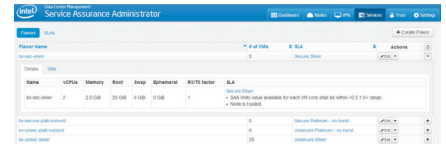

Abbildung 5: Intel Service Assurance Dashboard (Konfiguration von Flavors) ([17])

4.3 Konzeption zur SLA-Überwachung

Nach der Betrachtung der grundlegenden Elemente eines SLAs, der Problemstellung und Zielsetzung von Cloud-SLA-Management und der Analyse vorhandener Ansätze zum Umgang mit der Thematik, kann eine Architektur zur SLA-Überwachung in OpenStack konzipiert werden. Empfehlungen und vorhandene Ansätze der betrachteten Arbeitsgruppen sollen in die Konzeption einbezogen werden. Es soll auf vorhandene OpenStack Komponenten aufgebaut werden und eine integrierbare Erweiterung mit Möglichkeit zur Vorlage als Blueprint geschaffen werden.
Im ersten Schritt werden die vorhandenen Komponenten Ceilometer (Monitoring) und Heat (Orchestrierung) betrachtet, welche für die Umsetzung eines automatisierten SLA-Managements genutzt werden können.

4.3.1 Funktionsumfang von Ceilometer & Heat

Ceilometer erhält Messwerte und Events über den zentralen Nachrichtenbus (Oslo Notification Bus). Die meisten Services schicken Ihren Zustand und konkrete Werte bereits selbstständig auf den Bus, sodass ein ”Notification Agent” Änderungen liest und weiterverarbeitet. Nova Compute Nodes betreiben bspw. einen Resource Tracker, welcher standardmäßig alle 60 Sekunden die Ressourcenauslastung berichtet. [8] Für Werte, die nicht von den Modulen selbst publiziert werden, werden sog. ”Polling Agents” eingesetzt, welche explizit über die jeweiligen APIs benötigte Daten abfragen. Nachdem die Messwerte empfangen und falls notwendig normalisiert und transformiert wurden, werden diese in einer Datenbank mit zugehörigem Zeitstempel gespeichert. Die Messerte können über die Publishing Pipeline anderen Komponenten zur Verfügung gestellt oder explizit über eine REST-API abgefragt werden. [5] Die durch Ceilometer erhaltenen Werte (sog. ”Meter”) stellen lediglich einen Messpunkt (vgl. ETSI [10]) dar. Durch eine passende Transformation oder Kombination mit anderen Meters können erweiterte Metriken konzipiert werden.
Eine relativ neue Funktionalität von Ceilometer sind ”Telemtry Alarms” [4]. In aktuellsten Entwicklungen wurde entschieden diese Funktionalität in ein eigenes Projekt ”Aodh” auszulagern. Für die weitere Ausführung kann jedoch die bisherige Dokumentation durch Ceilometer herangezogen werden.

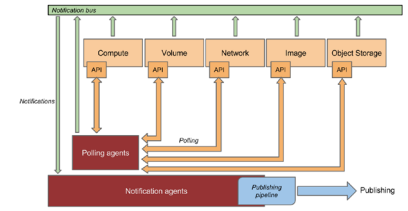

Abbildung 6: Ceilometer - Monitoring-Architektur (Datensammlung) ([4])

Die Alarmfunktion von Ceilometer unterstützt die Überwachung einzelner Messungen. Ein Alarm kann über mehrere Parameter definiert werden. Neben einer Bezeichnung und Beschreibung wird der zu verwendende Zähler (Meter), ein Grenzwert (Thresholds) mit Vergleichsoperatur, die ggf. anzuwendende Statistikfunktion und das zu überwachende Zeitfenster definiert. Beim Auslösen eines Alarms kann eine definierte Aktion gestartet werden bspw. ein ”Webhook” welcher die Alarmdaten an eine Webschnittstel-

le mitteilt. Der dargestellte Alarm wird ausgelöst, sobald die CPU-Auslastung einer Instanz 70% über die letzten 30 Minuten (10 Minuten Blöcke) überschreitet.

Listing 1: Exemplarischer Alarm in Ceilometer

```
ceilometer alarm-threshold-create
  --name cpu_hi \
  --description 'instance running hot' \
  --meter-name cpu_util --threshold 70.0 \
  --comparison-operator gt --statistic avg \
  --period 600 --evaluation-periods 3 \
  --alarm-action 'log://' \
  --query resource_id=INSTANCE_ID
```

Zusätzlich kann die Dimensionierung des Alarms eingeschränkt werden, sodass nicht der gesamte Ressourcenpool des Benutzers überwacht wird, sondern bspw. nur eine konkrete Instanz. Alle Zustandsänderungen und detaillierte Metadaten eines Alarms werden auch über dessen Deaktivierung oder Löschung hinaus gespeichert. Damit ist es möglich eine Langzeitanalyse bzw. Compliance-Berichte für den Kunden zu erstellen.
Ceilometer Telemetry Alarms können in Verbindung mit passenden Aktionssätzen (Policies) genutzt werden, um SLA-Metriken einzuhalten. Die konkrete Verwaltung (Erstellen, Aktualisieren, Löschen) von Alarm-Objekten muss jedoch von einer neuen Komponente übernommen werden.
Das Orchestrierungsmodul "Heat" nutzt sog. "Heat Orchestration Templates", welche eine Mensch- und maschinenlesbare Skriptsprache zum Aufbau von Cloudanwendungen ist. [4][Heat documentation] Heat setzt auf das Konzept von "Stacks". Ein Stack wird mittels eines Templates instanziiert und stellt eine Reihe Ressourcen mit entsprechender Konfiguration dar.
Heat Templates unterstützen die Verwendung von Ceilometer Alarmobjekten über den Ressourcentyp `OS::Ceilometer::Alarms`. Sowohl ein Alarm selbst als auch entsprechende Policies als auszuführende Aktionen bei Alarmauslöung lassen sich in einem HOT-Skript definieren. [4][Heat Template Guide]
Auch nach der Instanziierung eines Stacks ist der Zugriff im laufenden Betrieb bspw. über die REST-API von Heat möglich. So können nachträglich Alarmobjekte und Policies hinzugefügt, aktualisiert oder gelöscht werden.
Generell kann die Erweiterbarkeit, Flexibilität und Konfigurierbarkeit von OpenStack als sehr gut eingestuft werden. Mittels sog. "API Pipelines" können in weitestgehend jedem Modul Zwischenschritte bei der Verarbeitung eingefügt werden. Als Beispiel kann der Nova-Scheduler um neue Filter oder Gewichtungsmethoden erweitert werden und damit die Platzierung von VMs bspw. in Abhängigkeit von externen Parametern gesteuert werden. [12] Die Betrachtung der OpenStack Komponenten zeigt, dass bereits einige Funktionen zur Umsetzung einer SLA-Management Komponente vorhanden sind. Ceilometer kann für ein Monitoring einfacher Messpunkte und die Alarmierung bei Überschreitung von Grenzwerten eingesetzt werden. Heat stellt bereits die Funktion zur Verfügung Regelsatz-basiert im Alarmfall zu reagieren.
Für ein vollwertiges SLA-Management des gesamten Lifecycles sind weiterhin folgende Komponenten notwendig:

- Komponente zur automatisierten Verhandlung von SLAs gemäß eines Standards bzw. auf Basis von Service Templates (WS-Agreement Negotiation)
- Komponente zur Definition und Überwachung komplexer oder aggregierter Metriken (KQIs) auf Basis einzelner KPIs
- Komponente zur strukturierten Berichterstattung (Compliance Reports) mit Schnittstellen für externe Drittanbieter (Auditierung)

Der folgende Abschnitt zeichnet eine mögliches Grundkonzept zur Umsetzung dieser Ziele mit OpenStack auf.

4.3.2 SLA-Management für OpenStack

Für eine Erweiterung von OpenStack um SLA-Management Funktionalität soll ein "SLA-Manager" eingeführt werden. Auf die Authentifizierung und Autorisierung durch Keystone wird in der weiteren Ausführung nicht eingegangen. Es wird vorausgesetzt, dass Aktionen von verschiedenen Benutzern mit entsprechender Autorisierung ausgeführt werden. Die allgemeine Systemverwaltung kann bspw. nur von einem Administrator vorgenommen werden. Die Anforderung von neuen Ressourcen auf Basis eines SLAs oder die Berichterstattung erfolgt in Abhängigkeit der Berechtigungen eines Kunden bzw. dessen Ressourcenpools.
Der SLA-Manager erfüllt, in drei Unterkomponenten gegliedert, folgende Aufgaben.
Der "SLA-Agent" übernimmt die Aufgaben der Bereitstellung eines Service Catalogs, passender Service Level Templates und die Abwicklung des Verhandlungsprozesses mit dem Kunden. Der Dokumentenaustausch soll im XML-Format und auf Basis des WS-Agreements [15] erfolgen. Der Servicekatalog enthält verfügbare Angebote der OpenStack Architektur bspw. Nova-Compute Instanzen oder "Trove" (Database-as-a-Service) Optionen. Der Kunde nutzt ein zur Verfügung gestelltes Service Level Template und füllt dieses entsprechend seiner Serviceanforderungen aus. Neben allgemeinen Angaben zum Ressourcenbedarf können explizit zu verwendende Metriken (KQIs) mit Grenzwerten (Thresholds) und der Gewichtung untereinander angegeben werden. Verfügbare Metriken (in Abhängigkeit vom Monitoring) werden vom Service Provider verwaltet. Sollte eine Kundenanforderung nicht im aktuellen Service Level Template enthalten sein, kann diese über neu definierte Metriken und Parameter vom Provider ergänzt werden. Dies ist ein manueller Prozess, da nicht pauschal alle möglichen Kundenanforderungen bzw. industriespezifische Metriken von Beginn an abgedeckt werden können. Durch ständige Erweiterung des Service Level Templates um neue Metriken vervollständigt sich der Katalog selbstständig. Der genaue Prozess zur Verwaltung von SLA-Metriken wird im Kontext des "Metric Agents" erklärt.
Für die Verhandlung eines konkreten SLAs wird der WS-Agreement Negotation Standard [16] eingesetzt. Entsprechende Schnittstellen werden analog zu anderen OpenStack Komponenten als REST-API umgesetzt. Die bereitgestellten Methoden können sich bspw. am TMF617 SES Management Interface [20] orientieren, sodass die neue Konfigurationswünsche oder die aktuelle Konfiguration über die Abfragen `get/-setServiceConfiguration()` ausgeführt werden können. Nachdem der SLA-Agent eine Anfrage (Request) eines Kunden erhalten hat, wird geprüft, ob die gewünschten Anforderungen umgesetzt werden können. Dazu werden alle beteiligten Komponenten (bspw. Nova für Compute-Ressourcen, Neutron für Netzwerkkomponenten, etc.) befragt ob entsprechende Ressourcen verfügbar sind. Der SLA-Reporter sollte zur aktuellen Auslastung oder der Performance vergleichbarer Anwendungen befragt werden, um eine fundierte Evaluation erstellen zu können, ob der angefragte Service Level voraussichtlich geliefert werden kann. Sollte keine Neuverhandlung des SLAs notwendig werden, erstellt der SLA-Agent ein gemäß der Anfrage parametrisiertes Heat-Orchestration-Template und instanziiert den Stack. Der Metric Agent wird über den laufenden Stack und das zugehörige SLA informiert. Das SLA selbst wird im XML-Format mit Metadaten wie einem Zeitstempel und Kundeninformation in einer Datenbank zum späteren Abgleich und aus Dokumentationsgründen abgelegt.
Der "Metric-Agent" soll die Funktionalität anbieten komplexere und aggregierte Metriken (KQIs) auf Basis einfacher Ceilometer Messpunkte (Meter) zu definieren und zu überwachen. Zur Sicherstellung des Monitorings werden die vorhandenen Messpunkte von Ceilometer in Verbindung mit Telemetry Alarms eingesetzt. Wie zuvor dargestellt

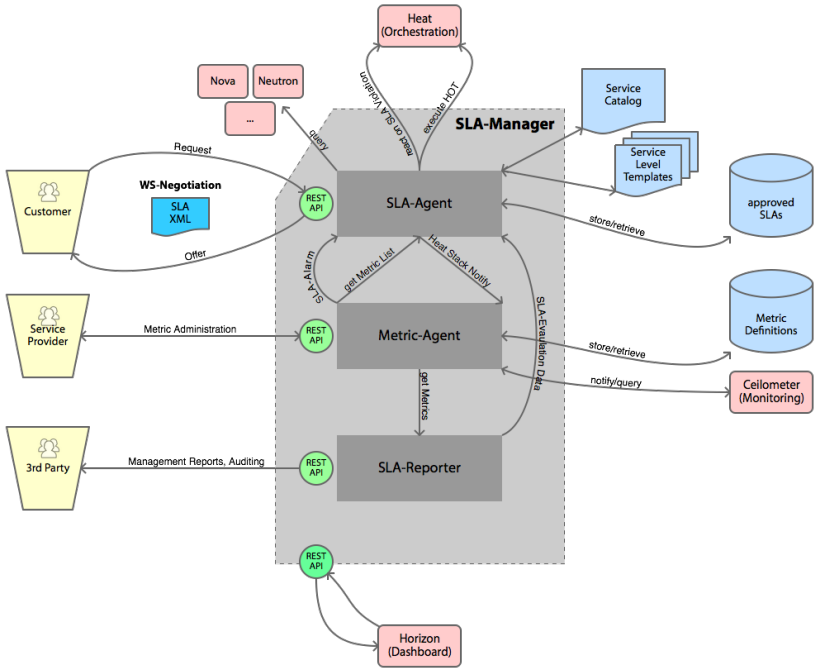

Abbildung 7: Konzept - SLA-Management Architektur für OpenStack

werden Quality of Service oder aussagekräftige Qualitätsmerkmale (Key Quality Indikatoren) durch Aggregation verschiedener Metriken oder Messungen erzeugt. Der Metric-Agent soll es dem Provider auf Basis von Best-Practices oder durch Vorschläge des Kunden ermöglichen individuelle Metriken zu definieren. Durch die Kombination vorhandener Messpunkte (Ceilometer Meters), anderer Metriken und Variablen mit statistischen Funktionen können komplexe Performancemetriken ausgedrückt werden. Metrikdefinitionen werden zusätzlich mit Metadaten wie der Klassifizierung (Speed, Accuracy, Reliability) und Kategorisierung (Businessmetrik für Billing oder technische Metrik für Performance) in einer Datenbank gespeichert. Sobald der Metric-Agent über die Instanziierung eines Heat-Stacks informiert wurde, werden die entsprechenden Ressourcen in Verbindung mit Ceilometer überwacht. Ein zusätzliches Speichern von Messwerten ist nicht notwendig, da entsprechende Daten von Ceilometer und der zum Zeitpunkt gültige SLA bereits abgelegt wurden. Somit ist auch eine spätere Überprüfung von Metriken ohne zusätzlichen Datenstand möglich. Der Metric-Agent überwacht kontinuierlich die definierten Metriken aktiver Stacks in Verbindung mit Ceilometer. Sollte sich eine negative Entwicklung bspw. mittels einer Trendanalyse abzeichnen oder es bereits zu einer Überschreitung von SLA-Thresholds gekommen sein, wird der SLA-Agent darüber informiert welcher entsprechende Aktionen in Form einer Orchestrierung mittels Heat vornimmt.

Der "SLA-Reporter" beschäftigt sich mit der Berichterstattung und Zurverfügungstellung von Monitoring-Daten an externe Nutzer (bspw. Drittanbieter welche die Auditierung des Providers vornehmen). Der Reporter dient der Aufbereitung vorhandener Monitoring-Daten in eine strukturierte und standardisierte Form. Das genaue Format wie Angaben an einen Nutzer der Schnittstellen weiterzugeben sind, kann in Verhandlungen mit dem Kunden definiert werden. Auch hier empfiehlt sich die Umsetzung einer REST-API und Schnittstellen bspw. auf Basis des TMF617 SES Management Interface [20], sodass eine Methode `getManagementReport()` Informationen über die Servicekonfiguration, den Zustand, aufgetretene Fehler und gemessene Metriken in einem XML-Dokument berichtet. Ähnlich der Ceilometer-Schnittstellen soll auch der SLA-Reporter parametrisierte Abfragen ermöglichen, sodass bspw. aggregierte Performancetrends eines bestimmten Kunden über einen gewissen Zeitraum abrufbar sind. Der Reporter soll vom SLA-Agent für die SLA-Evaluation eingesetzt werden, sodass ein Vergleich mit ähnlichen Auslastungsszenarien oder bereits erfüllter SLA-Konfigurationen möglich sein soll.

Neben den angebotenen REST-APIs soll ein weiteres Dashboard in der Weboberfläche zur Verfügung stehen. Sofern die beschriebenen Komponenten in die OpenStack Architektur integriert wurden, kann die Oberfläche von Horizon erweitert werden. [4][Building a Dashboard using Horizon]

5 Erkenntnisse & Ausblick

Die Betrachtung verschiedener Ansätze zur Umsetzung von SLA-Management in Cloud-Umgebungen hat gezeigt, dass eine deutliche Notwendigkeit an Standardisierung und strukturierten Prozessen zum Umgang mit SLAs und darin enthaltenen Performance-Metriken besteht. Die einzelnen Arbeitsgruppen empfehlen die Verfolgung eines gemeinsamen Standards durch Zusammenarbeit und Übernahme von erwiesenen "Best Practices". TMForum schlägt eine Standardisierung mittels Service Templates und Policy-basierten Cloud SLA-Management vor. Dies beinhaltet einen einheitlichen Satz an Basisvereinbarungen und Metriken, welche ggf. Industrieabhängig erweitert werden können. [20] Die SLA-Management Arbeitsgruppe von TMForum bietet bereits einen detaillierten Vorschlag zur Standardisierung und wird unter anderem durch die Verwendung des OGF WS-Agreement und des zugehörigen Negotiation-Prozesses als am vielversprechendsten gewertet. [12]

Die Problematik von Cloud-SLAs und gewünschte Ziele zeigen, dass besonders Handlungsbedarf im Bereich von Enterprise-SLAs und entsprechenden Performancegarantien für kritische Anwendungen besteht. Neben dem Kundeninteresse von Unternehmen, kritische Anwendungen zuverlässig mit Cloud-Infrastruktur betreiben, haben auch die Anbieter ein Interesse an fortgeschrittenen SLA-Werkzeugen. Ein verbessertes SLA-Management bewirkt eine genauere Kapazitätsplanung und damit eine Kostenersparnis, da die Gesamt-Cloud-Kapazität geringer angelegt werden kann. Weitere Vorteile versprechen sich Kunden und Anbieter aus der Möglichkeit proaktiv Handeln zu können. Können voraussichtliche Verstöße gegen SLA-Parameter frühzeitig erkannt werden, ist eine intelligente Migration von Ressourcen möglich. [9]

Der Einsatz von OpenStack zur Bereitstellung von Infrastruktur als Cloud-Service wird stets beliebter. Die Anwender erhoffen sich durch die offene Architektur und die hohe Flexibilität und Erweiterbarkeit der Software ein stabiles und kostengünstiges System, welches die individuellen Anforderungen abbilden kann. Es wird damit einem durch große IT-Firmen gesetztem Trend gefolgt. OpenStack bietet eine Vielzahl an APIs, sollte jedoch nicht als Management Tool sondern mehr als Framework gesehen werden. Eine vollwertige Produktivinstallation benötigt viele Arbeitskräfte und einige

Zeilen an Individualcode, weswegen Private Clouds mit OpenStack von der Presse derzeit noch mehr als "wissenschaftliche Projekte" [7] angesehen werden.
Zur Konzeption einer SLA-Management Komponente für OpenStack wurden die Module "Ceilometer" und "Heat" genauer betrachtet. In der aktuellen Version (v11 Kilo - Mai 2015) sind bereits einige Funktionalitäten zur Überwachung einzelner Messpunkte (Ceilometer Meters mit Telemetry Alarms) und zur entsprechenden Reaktion bei Verstößen gegen Vereinbarungen (Heat Orchestration Templates) vorhanden. Eine Lücke zwischen vorgestellten Bemühungen zur Standardisierung und den vorhandenen Möglichkeiten in OpenStack ergibt sich aus der fehlenden automatisierten Verarbeitung von SLAs und die Umsetzung entsprechend komplexer Metriken (KQIs), welche Qualitätsstandards anstelle konkreter Ressourcenwerte ausdrücken. Für die entsprechende Funktionalität muss eine Erweiterung, wie der vorgestellte "SLA-Manager", eingebunden werden.
Neben der Umsetzung und Validierung des vorgeschlagenen Konzepts lassen sich im Ausblick folgende Empfehlungen geben. Die vorgeschlagenen Ansätze von TMForum oder anderen Arbeitsgruppen sind eine solide Grundlage für weitere Forschungen. Die Arbeiten werden als gemeinnützig vorgestellt, dennoch werden die Gruppierungen jedoch durch entsprechende Industriebeteiligungen geführt, sodass in vielen Fällen ein wirtschaftliches Interesse im Vordergrund steht. Die von Intel präsentierte "Service Assurance Technology" [17] bietet derzeit eines der am weitesten fortgeschrittensten Entwicklungen. Die genutzte Technology der "Deep Telemetry Cache Inspection" beruht jedoch auf dem proprietären CPU-Design von Intel, sodass ein Interesse der Verkaufssteigerung bestimmter Hardware nahe liegt.
Nichtsdestotrotz sind die aktuellen Arbeiten als positiv zu bewerten, da alle Ansätze einen Einblick in mögliche Umsetzungen geben und die besten Vorgehensweisen extrahiert werden können. Die vorgestellte Architektur versucht die extrahierten Empfehlungen anzuwenden. Die konkrete Umsetzung benötigt jedoch eine komplexere Intelligenz in den Komponenten zur SLA-Evaluation und für die proaktive Reaktion. Weitere Forschungen in den Feldern der SLA-Evaluation basierend auf Statistiken und ähnlichen Szenarien, sowie Trendanalysen zur frühzeitigen Erkennung von Problemstellen sind erforderlich.

Literatur

[1] Wired - service level agreements in the cloud: Who cares? Website, 2011. http://www.wired.com/insights/2011/12/service-level-agreements-in-the-cloud-who-cares; Letzter Zugriff: 15.10.2015.

[2] IBM developerWorks - review and summary of cloud service level agreements aus 'cloud computing use cases whitepaper' version 4.0. Website, 2013. http://www.ibm.com/developerworks/cloud/library/cl-rev2sla.html; Letzter Zugriff: 11.06.2015.

[3] Cloudify - OpenStack Wiki in short – a quick guide to open cloud. Website, 2014. http://getcloudify.org/2014/07/18/openstack-wiki-open-cloud.html; Letzter Zugriff: 20.06.2015.

[4] OpenStack Dokumentation für Version 11 - Kilo (April 2015). Website, 2015. http://docs.openstack.org/kilo; Letzter Zugriff: 13.10.2015.

[5] OpenStack Wiki dokumentation. Website, 2015. https://wiki.openstack.org/wiki; Letzter Zugriff: 13.10.2015.

[6] The Register - HP snuggles up to OpenStack in cloud embrace. Website, 2015. http://www.theregister.co.uk/2015/08/07/openstack_hp_helion; Letzter Zugriff: 18.08.2015.

[7] The Register - OpenStack private clouds are science projects says gartner. Website, 2015. http://www.theregister.co.uk/2015/05/18/openstack_private_clouds_are_science_projects_says_gartner; Letzter Zugriff: 15.10.2015.

[8] Bauza, Sylvain and Dugger, Donald. OpenStack Summit 2015 Vancouver - Enhancing OpenStack Projects with Advanced SLA and Scheduling. Website, 2015. https://www.openstack.org/summit/vancouver-2015/summit-videos/presentation/enhancing-openstack-projects-with-advanced-sla-and-scheduling; Letzter Zugriff: 15.10.2015.

[9] Chadwick, Pete and Prakash, Alok. OpenStack Summit 2015 Vancouver - Pets vs Cattle - Meeting Enterprise SLAs in OpenStack. Website, 2015. https://www.openstack.org/summit/vancouver-2015/summit-videos/presentation/pets-vs-cattle-meeting-enterprise-slas-in-openstack; Letzter Zugriff: 15.10.2015.

[10] ETSI - Europoean Telecommunications Standards Institute. Network functions virtualisation (NFV); service quality metrics. Website, 2014. http://www.etsi.org/deliver/etsi_gs/NFV-INF/001_099/010/01.01.01_60/gs_NFV-INF010v010101p.pdf; Letzter Zugriff: 01.07.2015.

[11] B. S. Kaliski, Jr. and W. Pauley. Toward risk assessment as a service in cloud environments. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, pages 13–13, Berkeley, CA, USA, 2010. USENIX Association.

[12] C. A. Lee and A. F. Sill. A design space for dynamic service level agreements in openstack. *Journal of Cloud Computing*, 3(1), 2014.

[13] L. Leong. Cloud IaaS SLAs can be meaningless (gartner blog network). Website, 2012. http://blogs.gartner.com/lydia_leong/2012/12/05/cloud-iaas-slas-can-be-meaningless; Letzter Zugriff: 10.07.2015.

[14] C. Leymann, F. Fehling, R. Retter, W. Schupeck, and P. Arbitter. *Cloud Computing Patterns: Fundamentals to Design, Build, and Manage Cloud Applications*. Springer, Wien, 2014.

[15] Open Grid Forum. GFD192, Web Services Agreement Specification (WS-Agreement). Website, 20011. http://cloudindustryforum.org/downloads/standards/GFD.192.pdf; Letzter Zugriff: 15.10.2015.

[16] Open Grid Forum. GFD193, WS-Agreement Negotiation Version 1.0. Website, 2011. http://cloudindustryforum.org/downloads/standards/GFD.193.pdf; Letzter Zugriff: 15.10.2015.

[17] Redapt, Inc. and Intel Corporation. White Paper - Integrated OpenStack Cloud Solution with Service Assurance. Website, 2015. http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/intel-service-assurance-redapt-white-paper.pdf; Letzter Zugriff: 15.10.2015.

[18] M. Theoharidou, N. Tsalis, and D. Gritzalis. In cloud we trust: Risk-assessment-as-a-service. In C. Fernández-Gago, F. Martinelli, S. Pearson, and I. Agudo, editors, *Trust Management VII*, volume 401 of *IFIP Advances in Information and Communication Technology*, pages 100–110. Springer Berlin Heidelberg, 2013.

[19] TM Forum. GB917, SLA Management Handbook, Release 3.1 Best Practice. Website, 2012. https://www.tmforum.org/resources/standard/gb917-sla-management-handbook-release-3-1; Letzter Zugriff: 25.09.2015.

[20] TM Forum. TR178 Enabling End-to-End Cloud SLA Management V2.0.2 Standard. Website, 2012. https://www.tmforum.org/resources/technical-report-best-practice/tr178-enabling-end-to-end-cloud-sla-management-v2-0-2; Letzter Zugriff: 25.09.2015.