Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Masterthesis

Iwer Petersen

Kollaboration im virtuellen Team: Grenzen des Avatar-Realismus bei verteilter Echtzeit-Rekonstruktion

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Iwer Petersen

Kollaboration im virtuellen Team: Grenzen des Avatar-Realismus bei verteilter Echtzeit-Rekonstruktion

Masterthesis eingereicht im Rahmen der Master examination

im Studiengang Master of Science Computer Science
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Ing. Birgit Wendholt
Zweitgutachter: Prof. Dr.-Ing. Andreas Meisel

Eingereicht am: 3. August 2016

Iwer Petersen

Thema der Arbeit

Kollaboration im virtuellen Team: Grenzen des Avatar-Realismus bei verteilter Echtzeit-Rekonstruktion

Stichworte

3D Avatar, Echtzeit Oberflächenrekonstruktion, nonverbale Kommunikation

Kurzzusammenfassung

Die dreidimensionale Rekonstruktion von Menschen zur Darstellung in virtuellen Szenen ist in den letzten Jahren verstärkt weiterentwickelt worden. Sowohl die Rekonstruktion in Echtzeit zur Erstellung von dreidimensionalen Sequenzen als auch die hochqualitative Offline Rekonstruktion wurden bereits überzeugend realisiert. Zeit und Datenmenge stellen hierbei eine Herausforderung bei der Anwendung dieser Technologie auf verteilte Multi-User Szenarien dar. Diese Arbeit untersucht die Auswirkungen verschiedener Rekonstruktionsvarianten auf die Kommunikation des Avatars mit einem Menschen um einen Kompromiss zur Datenreduktion zu finden.

Iwer Petersen

Title of the paper

Collaboration in a virtual team: Limits of avatar realism with distributed realtime reconstruction

Keywords

3D avatar, real-time surface reconstruction, nonverbal communication

Abstract

Three dimensional surface reconstruction of people for the purpose of virtual representation has been developed increasingly in the last years. Both the reconstruction in real time to create three-dimensional sequences of 3D models as well as high quality offline reconstruction has been realized convincingly. The required computation time and the amount of data produced pose a challenge for the application of this technology to distributed multi-user scenarios. This paper examines the impact of several reconstruction methods on the communication between an avatar and a human in order to find a tradeoff for data-reduction.

Inhaltsverzeichnis

1 Einführung

In den letzten Jahren hat das Themengebiet Augmented- / Virtual-Reality (AR/VR) erneut verstärkte Aufmerksamkeit genossen. Angefeuert durch die Entwicklung innovativer Human-Machine-Interfaces (HMI) sowie die teils kreative, Zweck entfremdende Verwendung durch Bastler und Wissenschaftler, sind eine Vielzahl an Konzepten, Ideen und wissenschaftlichen Arbeiten erschienen. Die allgemeinen Fortschritte in der Computertechnik lassen zum Teil früher verworfene Ideen wieder machbar erscheinen und ermöglichen eine erneute Untersuchung auf Machbar- und Wirksamkeit.

Die möglichen Anwendungsbereiche der entwickelten Technologien konzentrieren sich im Moment auf den Unterhaltungsbereich. Vor allem im Bereich der Computerspiele wird, leicht nachvollziehbar, großes Vermarktungspotential gesehen. Aber auch im professionellen Umfeld tauchen immer mehr Anwendungsmöglichkeiten auf. Das Spektrum reicht hier von einfacheren Szenarien wie Produktpräsentation [1, 2, 3] über virtuelle Overlays zur Assistenz in der Medizin oder in technischen Montageszenarien [4] bis hin zu verteilten, kollaborativen Anwendungen [5, 6].

Bei kollaborativen Anwendungen in einer virtuellen Umgebung wird gerne von Tele-Immersion gesprochen. Gemeint ist damit, dass Personen an voneinander entfernten Orten in einer virtuellen Umgebung möglichst natürlich miteinander kommunizieren und mit der virtuellen Umgebung interagieren können. Während der Begriff der Immersion zum Teil umstritten war, mehren sich die Hinweise, dass dieser Effekt existiert. So haben zum Beispiel die Hirnforscher Pavone et al. [7] in einem VR Experiment, bei dem Probanden eine missglückende Aktion eines Avatars aus Ego-Perspektive betrachten, nachgewiesen, dass im Gehirn Fehlererkennungsmarker aktiviert werden, welche für die Erkennung eigener Fehler bekannt sind. Auch in der Psychologie interessiert man sich mehr und mehr für immersive Technologien, um zum Beispiel Phänomene wie Body-Ownership zu erforschen [8].

Die starke Fokussierung auf dreidimensionale AR-/VR-Umgebungen, die durch Entwicklungen im HMI Bereich sowie aus der Computergrafik ermöglicht wird, lässt Versuche, Telepräsenz durch zweidimensionale Videostreams zu realisieren in den Hintergrund treten. In einem Szenario mit mehr als zwei Personen ermöglicht eine dreidimensionale Darstellung beispielsweise

eine intuitivere Kommunikation, indem eine angesprochene Person schon durch die räumliche Zuwendung des Sprechenden zu ihm erkennen kann, dass er angesprochen wird. Um eine solche Bandbreite in der Kommunikation in einer virtuellen Umgebung zu ermöglichen, scheint sich ein zweidimensionaler Videostream schlecht zu eignen.

Die dreidimensionale Repräsentation von Personen in einer virtuellen Umgebung ist bereits auf verschiedenen Wegen realisiert worden. In den ersten Ansätzen wurden aus geometrischen Primitiven zusammengesetzte Avatare verwendet um eine Person darzustellen, welche durch Motion-Capturing Verfahren animiert wurden. Ein etwas erweitertes Prinzip bedient sich gängiger Videospieltechnik indem ein geschlossenes Polygongitter anhand eines Skelettes verformt wird. Durch mit Hilfe von 3D Scannern erfassten 3D Modellen der realen Personen ist es dann möglich, eine recht naturgetreue Repräsentation in einer virtuellen Umgebung bereitzustellen. Bei diesen Verfahren, welche von der Genauigkeit und der Auflösung der Motion-Capturing Verfahren abhängig sind, wird auch von selbst-animierten Avataren gesprochen [9].

Mit steigender Computerleistung, sowie durch Entwicklungen in der 3D-Datenverarbeitung und im Bereich der 3D-Kameras wird eine weitere Methode zur Erzeugung dreidimensionaler Repräsentanten plausibel. Ausgehend vom 3D Scanning haben sich einige Versuche zur Echtzeitrekonstruktion von Personen als vielversprechend erwiesen. Echtzeitrekonstruktion bedeutet hier, dass einzelne, vollständige 3D-Modelle einer Person, so schnell hintereinander erzeugt werden, dass sie als sich kontinuierlich verändert wahrgenommen werden können.

2 Motivation

Im Vorwort des Computers in Human Behavior Journals 2016 weisen Giard und Guitton [10] nicht nur auf die Auswirkungen von immersiver Technologie auf das Verständnis von Wahrnehmung hin. Durch die Ausdehnung unserer Sinne weit über körperliche Grenzen hinaus stellt demnach die Möglichkeit über immersive, vernetzte Systeme als Avatar an physikalisch weit entfernten Orten zu kommunizieren und zu interagieren unsere gängige Definition von Bewusstsein in Frage. Sie mahnen an, dass alleinige Fokussierung auf die Technologie sehr leicht vergessen lässt, dass Kommunikation eine Sache der menschlichen Interaktion, und nicht der Technologie ist.

In einer vielbeachteten Arbeit über "Embodied conversational interface agents" hat Cassell [11] die verschiedenen Ebenen der direkten Kommunikation zwischen Menschen herausgearbeitet und die menschliche Angepasstheit an diese Ebenen betont. Demnach gehören zu einer möglichst natürlichen Kommunikation neben der Sprache auch sogenannte nonverbale Kommunikationsformen. Vor allem erkennbare Mimik und Gestik fügen dem gesprochenen Wort zusätzliche Informationsebenen hinzu und unterstützen so die Kommunikation. Damit dies mit voneinander entfernten Personen möglich ist, müssen sie in ausreichender Realitätstreue in der virtuellen Umgebung dargestellt werden. In einer Untersuchung von Dodds et al. [9] hat sich gezeigt, dass Menschen weitaus besser mit einem animiertem Avatar als mit einem statischen Avatar kommunizieren können. Weitere in Abschnitt 3.3 beschriebene Arbeiten scheinen zu unterstützen, dass eine realistischere Darstellung eines Avatars die Kommunikation mit diesem vereinfacht und effektiver macht. Eine realistischere Darstellung erfordert in der Regel aber eine höhere Datenmenge, die wiederum sehr wahrscheinlich eine längere Verarbeitungs- sowie Übertragungszeit zur Folge hat. Mit der geforderten Rekonstruktionsrate ist die zur Verfügung stehende Zeit allerdings begrenzt und erfordert einen Kompromiss bei der Rekonstruktionsqualität. Hierzu muss die zu verarbeitende Datenmenge ausgehend vom optimalen 3D Modell so weit reduziert werden, dass die Verarbeitung und Übertragung innerhalb der zur Verfügung stehenden Zeit erfolgen kann. Dabei sollten aber Faktoren, die eine intuitivere Kommunikation ermöglichen, möglichst nicht beeinflusst werden.

Diese Arbeit hat das Ziel herauszufinden welche Faktoren eines in Echtzeit rekonstruierten Avatars Einfluss auf die Kommunikation zwischen dem Avatar und einem Menschen haben, um einen Kompromiss in der Rekonstruktionsqualität des Avatars zu finden welcher die Kommunikation möglichst minimal beeinflusst. Dazu werden in dieser Arbeit Sequenzen von rekonstruierten Avataren in verschiedenen Qualitäten erzeugt und in einem Kommunikationsspiel gegen die Mensch zu Mensch Kommunikation evaluiert. Testpersonen versuchen dabei, von einem Avatar pantomimisch beschriebene Begriffe zu erraten.

Im Folgenden wird zunächst in Kapitel 3 ein Einblick in die Grundlagen der verschiedenen verwendeten Technologien gegeben, auf einige aktuelle Entwicklungen im Bereich der Echtzeit-Rekonstruktion eingegangen sowie ein Überblick über ähnliche Evaluierungen gegeben welche darauf hinweisen, dass eine realitätsgetreuere Darstellung in der Regel die Kommunikation verbessert. In Kapitel 4 werden die Anforderungen an die zu entwickelnde Software definiert. Anschließend wird in Kapitel 5 die entwickelte Lösung beschrieben. Diese umfasst Komponenten zur Kalibrierung der Sensoren, zur Aufnahme von Sensordaten sowie zur Verarbeitung der Sensordaten zu Sequenzen von 3D Modellen. Mit Hilfe der entwickelten Lösung werden in Kapitel 6 vier unterschiedliche Rekonstruktionsvarianten in einem Kommunikationsspiel auf die Auswirkungen auf die Kommunikation untersucht. Dabei werden Thesen aus früheren Untersuchungen zur Kommunikationsfähigkeit von Avataren bestätigt. Es zeigt sich aber auch, dass sich aufwändigere Rekonstruktionsverfahren durchaus lohnen können.

3 Grundlagen und Vergleichbare Arbeiten

Um einen Überblick in den aktuellen Stand der Forschung zu bekommen werden in diesem Kapitel zunächst in Abschnitt 3.1 grundlegende Technologien beschrieben, die ganz unterschiedliche Implikationen auf Verarbeitungszeit sowie Rekonstruktionsqualität haben. In Abschnitt 3.2 werden dann verschiedene Ansätze zur dreidimensionalen Rekonstruktion von Menschen in dynamischen Umgebungen betrachtet. Dabei wird deutlich, dass sowohl die Online-Rekonstruktion einzelner Frames in Echtzeit als auch die nahezu fotorealistische Offline-Rekonstruktion technisch machbar sind. In Abschnitt 3.3 werden dann einige Arbeiten betrachtet, die sich mit der Kommunikation mit virtuellen Avataren auseinandergesetzt haben und prinzipiell eine höhere Avatarqualität mit einer besseren Kommunikation in Verbindung bringen. Allerdings stellen sie auch fest, dass Avatare höherer Qualität anfälliger für kleine Imperfektionen sind.

3.1 Grundlagen

Durch die Verfügbarkeit neuer günstiger 3D-Sensorik sind in den letzten Jahren einige Forschungsarbeiten zur 3D Rekonstruktion erschienen. Hierbei werden verschiedene Sensoren und Verfahren eingesetzt, um Technologien aus dem Bereich 3D-Scanning Echtzeit-fähig zu gestalten. Die Zielsetzung reicht hier von Realisierung eines verteilten Tele-Immersions-System über die Erstellung statischer Szenen durch Echtzeit-Integration zeitlich aufeinander folgender Daten, bis hin zur Erstellung von hochqualitativen dreidimensionalen Free-Viewpoint-Videos .

3.1.1 Sensorik

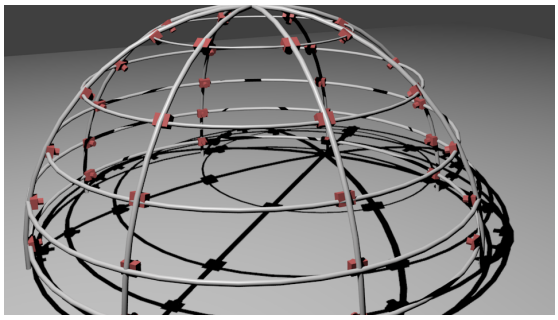


Abbildung 3.1: Multi-Kamera Array Beispiel-Setup: Halbkugel-Dome mit vielen RGB- und Infrarot-Kameras (rot). Quelle: eigenes Werk.

Die eingesetzten Technologien unterscheiden sich je nach Anwendungsfall in unterschiedlichen Bereichen. Als Datenquellen kommen in der Regel RGB- und Infrarotkamera-basierte Systeme zum Einsatz. Während die RGB Kameras die Daten für die Texturierung liefern, werden aus den Infrarotbildern nach dem Prinzip der Photogrammetrie Tiefeninformationen generiert. Dazu muss für jede Kamera die räumliche, sowie die Linsen-Kalibrierung bestimmt werden. Zu einem Multi-Kamera-Array zusammengeschlossen und, zum Beispiel wie in Abbildung 3.1 dargestellt, in Halbkugelform um ein Volumen montiert, kann eine Szene aus allen Richtungen aufgenommen werden. Ein solches System ist sehr flexibel durch die Möglichkeit die Kamerakonfiguration frei zu ändern und kann sehr hoch aufgelöste Ergebnisse erzielen. Allerdings ist die zu verarbeitende Datenmenge in der Regel ein Hindernis für eine Verarbeitung in Echtzeit.

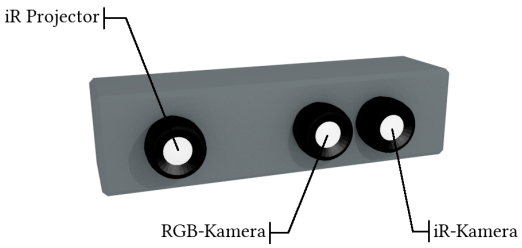


Abbildung 3.2: RGB-D Kamera, feste Systemkalibrierung durch gemeinsame Montage. Quelle: eigenes Werk.

RGB-D Kameras stellen insofern eine Besonderheit dar, als dass sie ein kalibriertes Infrarot Stereokamerapaar mir einer ebenso kalibrierten RGB Kamera in einem gemeinsamen Gehäuse

darstellen. Das Stereokamerapaar wird dabei aus einer Infrarotkamera und einem Infrarotprojektor aufgebaut, welcher nach den Gesetzen der Optik als invertierte Kamera verstanden werden kann. RGB-D Kameras generieren dann Tiefeninformationen aus der Verzerrung des bekannten, vom Projektor ausgestrahlten Musters wenn es von der Infrarotkamera aufgefangen wird. Durch die feste Konfiguration der Kameras kann dieser Schritt durch spezielle Hardware unterstützt werden, so dass direkt RGB- und Tiefenbilder ausgeliefert werden können. Mit vergleichsweise geringen Aufwand lassen sich die Pixel eines Tiefenbildes mit Hilfe der Kameraintrinsik auf einen 3D Punkt projizieren und so dreidimensionale Punktwolken erzeugen.

3.1.2 Polygongitter

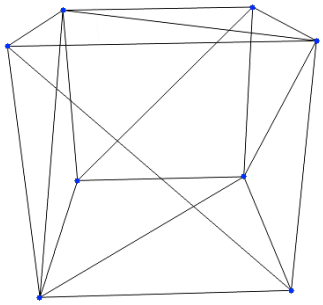


Abbildung 3.3: Würfel als Dreiecks-Polygongitter in Wireframe-Darstellung. Vertizen (blau) bilden je zu dritt Dreiecke. Quelle: eigenes Werk.

Die klassische Datenstruktur zur Darstellung eines 3D Modells ist das Polygongitter. Ein Polygongitter setzt sich aus einer Menge an indizierten Vertizen zusammen, welche je nach Art der Polygone mit drei oder mehr Vertizen Flächen bilden. Diese Struktur ist sehr gut verstanden und kann als de-facto Standard in allen gängigen 3D Verarbeitungs- und Darstellungstechnologien gefunden werden. Es existieren umfangreiche Werkzeuge zur Verarbeitung von Polygongittern sowie zur Triangulation von Polygongittern aus Punktdaten. Somit ist die Rekonstruktion eines einzelnen Frames direkt aus den aus Tiefenbildern generierten Punktwolken realisierbar.

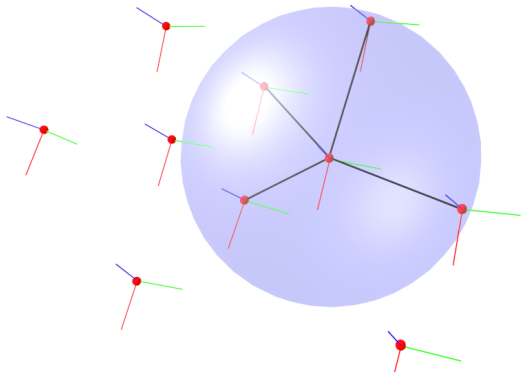


Abbildung 3.4: 3D Nachbarschaftssuche, Nachbarpunkte kommen als Triangulationspartner in Frage. Quelle: eigenes Werk.

Die Triangulation eines dreidimensionalen Polygongitters in Echtzeit ist zunächst eine große Herausforderung. Für jeden Punkt muss, wie in Abbildung 3.4 dargestellt, die lokale Nachbarschaft nach Punktkandidaten abgesucht werden, welche durch eine Kante verbunden werden sollen, um so nach und nach Polygone zu bilden. Ist bereits eine strukturelle Nachbarschaftsbeziehung bekannt, zum Beispiel bei der Generierung von Punktwolken aus 2D Bildern, kann die Suche entfallen und die Triangulation so erheblich beschleunigt werden (siehe Holz und Behnke [12]). Da die Nachbarschaftsbeziehungen aber nur im Kameraraum gelten, kann je Kamera lediglich ein partielles Modell der sichtbaren Oberfläche rekonstruiert werden.

Ein anderes Verfahren welches nah an die Echtzeitanforderung herankommt vereinfacht die dreidimensionale Nachbarschaftssuche indem um einen Punkt eine lokale Ebene aufgespannt wird, auf die die benachbarten Punkte projiziert werden. In dieser Ebene kann die Nachbarschaftsbeziehung durch eine zweidimensionale Suche festgestellt werden können. Diese Greedy-Projection-Triangulation von Marton et al. [13] erlaubt die Triangulation einer Punktwolke, die aus den Daten mehrerer Kameras zusammengesetzt wurde und somit keine eindeutigen Nachbarschaftbeziehung aus den Tiefenbildern mehr beinhaltet.

3.1.3 Distanzfunktion

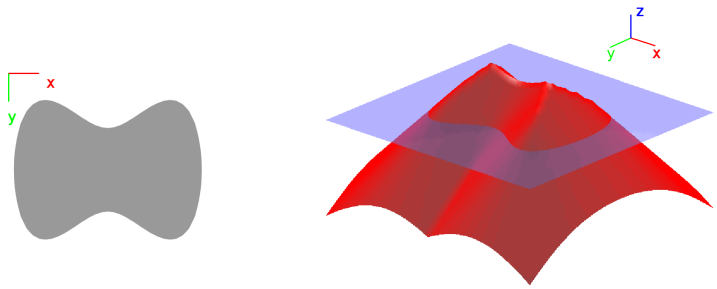


Abbildung 3.5: Zweidimensionale Form (links) und ihre Distanzfunktion (rechts, rot). Quelle: angelehnt an [14]

Ein anderer Anwendungsfall ist die temporale Integration zeitlich aufeinanderfolgender Frames in ein statisches 3D Modell (siehe [15]). Hierfür bietet sich eine volumetrische Datenstruktur an. Eine solche Datenstruktur ist die signierte Distanzfunktion (siehe [16]). Eine Distanzfunktion bildet auf jeden Punkt in einem Volumen seine kürzeste Distanz zu einer Oberfläche ab. Dabei werden Werte vor der Oberfläche positiv signiert, während Werte hinter der Oberfläche mit negativem Vorzeichen belegt werden. Die Oberfläche ist dann implizit durch den Null-Durchgang der Funktion definiert. In Abbildung 3.5 ist eine Distanzfunktion einer nichttrivialen, zweidimensionalen Form dargestellt. In der Abbildung sind die Werte der Distanzfunktion in rot auf die Z-Achse abgebildet. Der Null-Durchgang durch die blaue Ebene bildet die nebenstehende Form ab. Die Werte einer Distanzfunktion im dreidimensionalen Raum werden zum Beispiel in einem Octree gespeichert (siehe [17]). Durch Tracking der Kameraposition durch eine Iterative-Closest-Point Transformationabschätzung (siehe [18]) zwischen aufeinanderfolgenden Frames können die Tiefenwerte in den Octree integriert werden, indem der Octree in die Kameraperspektive transformiert wird, so dass die Tiefenwerte entlang der Z-Achse der Kamera in den Octree projiziert werden können. So kann das rekonstruierte Modell nach und nach aus verschiedenen Perspektiven vervollständigt werden. Auch Ungenauigkeiten wie durch Bildrauschen verursachte Rauheit der rekonstruierten Oberfläche werden so weitgehend ausgeglichen. Eine Echtzeit-Integration von Tiefendaten in eine Distanzfunktionen ist bisher lediglich durch GPU-beschleunigte Berechnung realisiert worden. Die Distanzfunktion wird dabei vereinfacht gesagt als dreidimensionales Array im GPU Speicher gehalten, während die Integration der Tiefenbilder über die Textur-Pipeline der GPU realisiert wird. Die Größe des Grafikspeichers legt somit eine harte obere Schranke für das Produkt aus der Größe des abgebildeten Volumens und der räumlichen Auflösung der Distanzfunktion fest. Will man also den

Detailgrad der Geometrie eines rekonstruierten 3D Modells erhöhen, muss man automatisch das erfassbare Volumen reduzieren. In der Praxis stellt das Tracking der Kameraposition durch Bestimmung der Frame-zu-Frame Transformation der Eingangspunktwolken eine gewisse Schwierigkeit dar. Die Zuverlässigkeit des ICP Verfahrens hängt stark von der geometrischen Beschaffenheit der zu scannenden Oberfläche und der Art der Feature-Deskriptoren, über die Korrespondenzen zwischen zwei Frames bestimmt werden, ab. Werden nicht genügend oder falsche Korrespondenzen gefunden, werden die Tiefenwerte aus einer falschen Perspektive auf das Model projiziert. Weiterhin verursacht das vorhandene Sensorrauschen über die Zeit eine gewisse Glättung feiner Strukturen. In der Praxis bedarf es einer ruhigen Hand und sachter Bewegungen um mit dieser Methode ein guten 3D Scan zu erhalten.

3.1.4 Texturierung

Um ein trianguliertes Polygongitter, welches zunächst nur geometrische Informationen beinhaltet, farbig darzustellen sind weitere Informationen nötig. Bei der Darstellung von Polygongittern wird zwischen Vertexfarben und Pixelfarben unterschieden. Mit Vertexfarben erhält jeder Punkt einen RGB Farbtupel während mit Pixelfarben in der Regel eine Textur mit übergeben werden muss.

Vertex-Farben

(a) (b) (c)

Abbildung 3.6: Vertexfarben auf Mesh mit hoher (a), mittlerer (b) und niedriger (c) geometrischer Auflösung. Quelle: eigenes Werk.

Vertexfarben werden pro Vertex festgelegt, und an die Grafikkarte übergeben. In der Rasterstufe der GPU werden diese Farben dann über die Polygone interpoliert. Jedes Polygon erhält so einen Farbverlauf aus den beteiligten Vertizen. Dadurch ist die erreichbare Darstellungsqualität, was die Texturierung angeht, direkt an die geometrische Auflösung der Polygongitters gekop-

pelt. In Abbildung 3.6 ist die Verringerung der Texturqualität mit sinkender geometrischer Polygonmesh Auflösung deutlich sichtbar.

Um diese Farben für jeden Vertex aus RGB Bildern zu bestimmen, muss die 3D Position des Vertex analog zum Lochkameramodell durch die Kameramatrix der RGB Kamera auf die zweidimensionale Bildebene projiziert werden. Durch Kenntnis der Linsenparameter (Intrinsik) sowie der Raumposition und -orientierung der Kamera (Extrinsik) wird eine perspektivische Projektionsmatrix konstruiert welche den 3D Punkt auf die 2D Bildebene abbildet. Im Falle der RGB-D Kameras sind in der Regel Methoden zur Abbildung der Farb- auf die Tiefeninformationen mitgeliefert. So kann die Vertexfarbe eines Punktes, welcher aus dem Tiefenbild Pixel $P_{depth}\binom{x}{y}$ generiert wurde, direkt aus dem RGB Pixel $P_{rgb}\binom{x}{y}$ ausgelesen werden.

Texturmapping

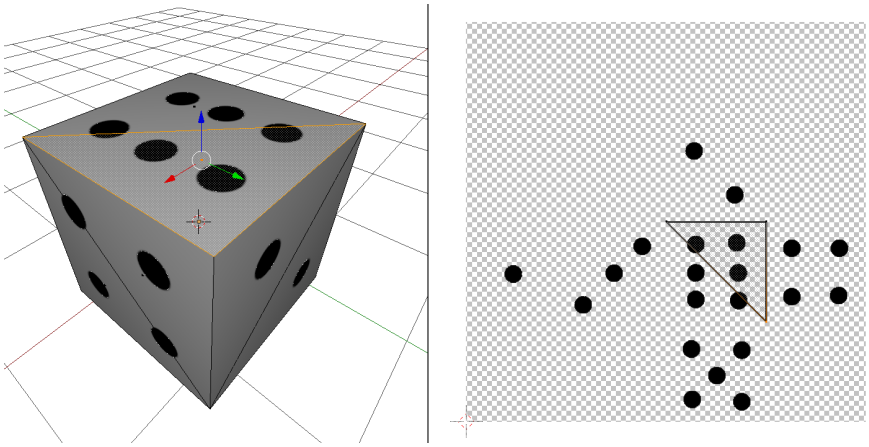


Abbildung 3.7: Texturmapping: Texturkoordinaten an jedem Vertex bestimmen den Texturausschnitt welcher auf ein Dreieck gerendert wird. Quelle: eigenes Werk

Um die mit einer RGB Kamera aufgenommene Textur per Pixel auf die Polygone zu rendern, wird statt einer Farbe pro Vertex gleich die gesamte Textur an die Grafikkarte übergeben. Das Polygongitter wird dann durch Texturkoordinaten für jeden Vertex erweitert. Nun werden in der Rasterstufe der GPU nicht Farben, sondern diese Texturkoordinaten interpoliert. Im Fragmentshader kann nun an den interpolierten Texturkoordinaten die Farbe aus der übergebenen Textur entnommen werden. So ist die Texturqualität des gerenderten Polygongitters nicht mehr von der geometrischen Auflösung abhängig, sondern von der Auflösung des Kamerabildes (siehe Abbildung 3.8).

Abbildung 3.8: Per Pixel Farben durch Texturemapping, deutlich bessere Qualität bei gleicher geometrischer Auflösung wie in Abbildung 3.6c. Quelle: eigenes Werk.

Ähnlich wie bei der Bestimmung der Vertexfarben kann zur Bestimmung der Texturkoordinate die 3D Position des jeweiligen Vertex auf die Bildebene der entsprechenden Kamera projiziert werden. An den Vertex wird dann die projizierte 2D Position angehängt.

Ohne Zweifel ist die Datengröße eines 3D-Modells mit Texturmapping durch die zusätzlich notwendige Textur um einiges größer. Die Übertragung von Bilderstreams mit den nötigen Codierungs- und Kompressionsmethoden ist jedoch ein gut erforschtes Feld.

3.2 Verschiedene Ansätze zur Rekonstruktion

Eine der frühesten vollständigen Implementierungen einer tele-immersiven Anwendung auf Basis von in Echtzeit erfassten dreidimensionalen Daten wird in der Arbeit von Mekuria et al. [19] beschrieben. Mit dem Rekonstruktionssystem von Alexiadis et al. [20], werden basierend auf mehreren Tiefenbildkameras laufend 3D Daten einer Person von allen Seiten erfasst. Aus den jeweiligen entstehenden Tiefenbildern wird die Person extrahiert, und ähnlich zu dem in Abschnitt 3.1 beschriebenen Verfahren mit Hilfe der aus den Bildern bekannten Nachbarschaftsbeziehungen in Polygongitter-Halbschalen umgewandelt. Durch Zusammenfügen der verschiedenen Halbschalen und Bereinigen der sich dann überlappenden Polygone entsteht so ein vollständiges Modell der Person. Durch einen ausgeklügeltes Kompressionsverfahren können die entstehenden Polygongitter dann über ein Netzwerk zum Beispiel an eine Visualisierungskomponente ausgeliefert werden.

Der Vorteil dieses Verfahrens ist die schnelle Triangulation des Polygongitters durch Verarbeitung der Punktwolken der einzelnen Kameras im Kamerakoordinatensystem. Durch die

bekannten Punktnachbarschaften kann so ein Großteil des Gitters in kurzer Zeit trianguliert werden. Nachteilig wirkt sich die Verwendung von Vertexfarben aus. Um mit dieser Art der Kolorierung eine hohe Auflösung der Textur zu erreichen muss die Punktdichte vergleichsweise hoch gehalten werden muss. In der Konsequenz fällt eine wesentlich größere Datenmenge an, die weiter verarbeitet werden muss.

Dieses Grundkonzept eines tele-immersiven Systems wurde von Pejsa et al. [21] durch einen Projektor erweitert um ein AR Telepräsenz-System zu implementieren. Hierbei werden Projektor RGB-D Kamera Einheiten (ProCam Units) verwendet, um das 3D Modell des lokalen Benutzers zu erfassen und das des entfernten Benutzers durch die Kopfposition des lokalen Benutzers perspektivisch vorverzerrt zu projizieren. Da das Szenario eine zugewandte Konversation zwischen zwei Personen abbildet wird hier lediglich ein partielles Modell aus einer frontalen Kamera erstellt. Das entwickelte System wurde mit einer kooperativen Aufgabe hinsichtlich Bearbeitungszeit, gefühlter Präsenz und Kommunikationsqualität verglichen. Die Aufgabe bestand in darin, aus Bauklötzen eine Form nach Anleitung zusammenzustellen. Dabei wurde die Anleitung von einem Teilnehmer gegeben, während die Bauklötze von einem anderen Teilnehmer zusammengestellt wurden. Die zur Durchführung der Aufgabe nötige Kommunikation wurde zum einen direkt, zum anderen über Videotelefonie und über das entwickelte tele-immersive System durchgeführt. Erwartungsgemäß ist in allen drei Kriterien Bearbeitungszeit, gefühlter Präsenz und Kommunikationsqualität die direkte Kommunikation den Telepräsenzsystemen überlegen. Es zeigt sich jedoch, dass die gefühlte Präsenz sowie die Bearbeitungszeit signifikant optimaler mit dem entwickelten System als mit der Videokonferenzsoftware bewertet wurden.

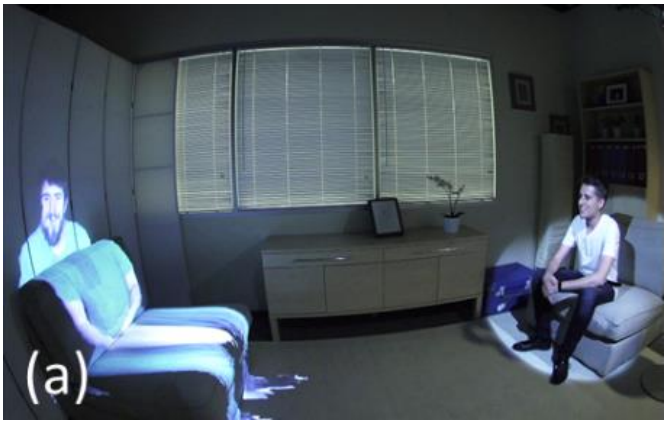


Abbildung 3.9: Room2Room Szenario: Projizierter Avatar entfernter Person und lokale Person im selben Raum. Quelle: [21].

Bei dieser Arbeit wird vor allem der Vorteil von immersiven Systemen gegenüber der Videotelefonie klar, was den Kommunikationsaspekt angeht. Die Beschränkung auf die frontale Ansicht einer einzelnen Kamera optimiert dieses Verfahren für ein Face-to-Face Szenario in ein projizierten AR Umgebung für einen einzelnen Benutzer pro Raum. Es ist aber weder für Multi-User Szenarien geeignet, da sich die Projektion eines entfernten Benutzers nur für einen lokalen Benutzer perspektivisch korrekt darstellen lassen. Da mit diesem Verfahren nur ein partielles, frontales 3D Modell erstellt wird, eignet sich das Verfahren auch kaum für Szenarien die mit AR/VR-Brillen arbeiten, in denen Benutzer auch aus Perspektiven gesehen werden können, die weit von der Kamera Position abweichen.

Im Bereich der Echtzeitrekonstruktion ist die Microsoft Research Gruppe Interactive 3D Technologies seit einigen Jahren sehr aktiv. Bereits 2011 präsentierten sie weitergehende Entwicklungen wie die Arbeit von Izadi et al. [22] die auf Integration temporal sukzessiver Frames in ein globales 3D Oberflächenmodell setzt. Die Arbeit präsentiert eine Echtzeit fähige Implementierung der in Abschnitt 3.1.3 beschriebenen signierten Distanz Funktion. Zunächst hatte dieses Verfahren den Makel, dass eine Veränderung der Szene während des Scans dazu führte, dass die nun veränderte Oberfläche im Model über einen Zeitraum von mehreren Frames verschwindet und an neuer Stelle wieder auftaucht. Dadurch war das Verfahren zunächst nur für statische Szenen zu gebrauchen.

Abbildung 3.10: DynamicFusion: Integration aufeinanderfolgender Frames unter dynamischen Bedingungen. Von links nach rechts wird das 3D Modell immer weiter vervollständigt. Quelle: [23].

Kürzlich wurde das Verfahren durch Newcombe et al. [23] um einen dynamischen Verformungstracker erweitert, welcher es ermöglicht auch dynamische Szenen nach obigen Verfahren über die Zeit zu erfassen und zugleich die aktuellen Verformungen auf das sich im Entstehen befindende, eigentlich statische 3D Modell zu übertragen. Ein kürzlich unter dem Titel "Holoportation"vorgestelltes Teaser-Video [6] präsentiert eine vollständige Pipeline zur Echtzeit 3D Darstellung entfernter Szenen und kündigte einige Paper im Zusammenhang mit dem Projekt an. Die Titel der im Laufe dieses Jahres zu veröffentlichenden Paper suggerieren signifikante Fortschritte im Bereich der Fusion Technologien [24], der Datenverarbeitung der RGB-D Sensoren [25] und der zur Netzwerkübertragung verwendeten Kompressionsmethoden [26].

Die temporale Integration der durch die Sensoren frameweise gelieferten Daten in ein einziges 3D Modell muss hier eindeutig als Vorteil deklariert werden. Trotz der Begrenzung der Auflösung oder des Aufnahmevolumens durch die Größe des GPU-Speichers bietet dieses Verfahren großes Potential für die Echtzeitrekonstruktion. Die absehbare Entwicklung im Bereich der Grafikhardware zeigt starkes Wachstum im Bereich der Speicherausstattung und lässt dieses Verfahren zukünftig interessanter werden. So ist diese Limitierung zur Zeit aber auch der größte Nachteil dieses Verfahrens.

Collet et al. [27] zeigen, mit welcher Qualität Sequenzen von 3D Modellen mit fotogrammetrischen Methoden erstellt werden können, wenn die Daten nicht in Echtzeit verarbeitet werden müssen. In dieser Arbeit wird eine Pipeline zur Erstellung von hoch aufgelösten Free-Viewpoint Videos vorgestellt, welche auf Basis der Daten von 106 RGB- und Ir-Kameras arbeitet, die wie in 3.1 beschrieben um ein Volumen herum angeordnet sind. Auch hier werden zunächst Punktwolken erzeugt, welche mit einer Poisson Rekonstruktionsvariante trianguliert werden. Das besondere an dieser Arbeit ist die Idee, anschließend die Meshauflösung an die Details der Topologie anzupassen. Dadurch werden Bereiche mit feineren topologischen Details mit mehr Vertizen und somit kleineren Polygonen abgebildet als Bereiche die eine eher flache Topologie darstellen. Weiterhin werden die 3D Modelle zwischen Keyframes, die direkt aus den Kameradaten rekonstruiert werden, anhand der Kameradaten verformt. Dadurch kann die Textur des 3D-Modell über mehrere Frames wiederverwendet werden. Wie oben erwähnt, arbeitet dieses System nicht in Echtzeit, sondern zielt vielmehr auf einen möglichst hohen geometrischen Detailgrad. Das Ziel bei diesem System ist, die resultierenden Modelle in einen MPEG Stream zu codieren um sie auf einem entfernten System in Echtzeit darstellen zu können.

Abbildung 3.11: Beispiele der nahezu fotorealistisch rekonstruierten Avatare nach dem Verfahren von Collet et. al. Quelle: [27].

In Bezug auf die vorliegende Arbeit ist der Vorteil dieses Verfahren die deutlich verbesserte Qualität der rekonstruierten Sequenzen von 3D Modellen. Die dynamische Anpassung der geometrischen Auflösung an die Struktur der Oberfläche dient auch der Optimierung

des Kompromisses zwischen Qualität und Datenmenge. Auch die Einsparung durch die Texturwiederverwendung ist auf die Reduktion des Datenaufkommens ausgelegt. So kann das Rekonstruktionsergebnis als hochqualitativer Stream bereitgestellt werden. Nachteil des Verfahrens ist die für die Rekonstruktion benötigte Rechenzeit, die sich nicht für ein Echtzeitsystem eignet.

3.3 Evaluierungen der Kommunikation zwischen Avatar und Mensch

Die Frage ob eine realistischere Darstellung eines Avatars überhaupt einen positiven Effekt auf die menschliche Perzeption ist schon älter als die aktuelle Beschäftigung mit VR und AR. Schon 1970 schrieb Mori [28] über den Effekt des “Uncanny-Valley“, welcher bei steigendem Realismusgrad der Darstellung eines Menschen den gegenteiligen Effekt bewirken, und unheimlich werden kann. Neuere Arbeiten aus dem Bereich der VR, welche einzelne Aspekte von detailierteren Avataren untersucht haben, weisen jedoch auf deutliche Verbesserungen in der Kommunikation eines Menschen mit einem Avatar hin.

Garau et al. [29] untersuchten verschiedene Avatardarstellungen und ihren Einfluss auf die Kommunikation in virtuellen Umgebungen. Hierzu wurden zwei Testkandidaten jeweils in eine Cave Augmented Virtual Environment (CAVE) und in ein Head-mounted Display Setup (HMD) gesteckt um dann mit dem virtuellen Avatar des anderen zu kommunizieren. Der Avatar wurde wie in Abbildung 3.12 dargestellt einmal unrealistisch als humanoides Strichmännchen ohne erkennbares Geschlecht dargestellt und einmal als näherungsweise fotorealistische Abbildung eines Mannes oder einer Frau. Weiterhin wurden die Augen des Avatars zum einen zufällig und zum anderen geleitet von der Stimme des Sprechenden animiert.

Abbildung 3.12: Avatardarstellung mit niedrigem Realismusgrad (links), und höherem Realismusgrad in männlicher (Mitte) und weiblicher Ausprägung (rechts). Quelle: [29].

Die Untersuchung weist darauf hin, dass unabhängig vom Realismusgrad des Avatars eine sinnvolle Animation der Blickrichtung zu einer stärkeren Präsenz in der virtuellen Umgebung führt. Durch den Eindruck eines hergestellten Augenkontaktes fühlt sich der Betrachter angesprochen und wahrgenommen. Weiterhin stellt die Untersuchung fest, dass ein realistischer Avatar die gefühlte Kommunikationsqualität signifikant verbessert, so lange der Avatar den Stimmen-geleiteten Augenkontakt herstellen kann. Bei Verwendung zufällig animierter Blickrichtung zeigen sich hier die negativen Effekte des Uncanny-Valleys. Die Arbeit schließt mit der Vermutung, dass mit steigendem Photorealismus auch die Anforderungen an ein realistisches Verhalten des Avatars steigen.

Abbildung 3.13: Links: Motion-Capturing zur Selbstanimation der Avatare. Rechts: virtuelle Untersuchungsumgebung. Quelle: [9].

Etwas weiter gehend beschäftigen Dodds et al. [9] sich mit der Gestik von Avataren. Sie weisen nach, dass durch die Möglichkeit mit Händen zu gestikulieren die Kommunikation mit selbst-animierten Avataren ausdauernder geführt wird als mit Avataren in statischer, neutraler Pose. Bei dem Versuch hatten je zwei Teilnehmer die Aufgabe, in einer VR Umgebung ein Wörterratespiel zu spielen. Die Avatare der Teilnehmer wurden zum einen mittels Motion-Tracking animiert (siehe Abbildung 3.13), zum anderen aber in neutralen Posen eingefroren, oder mit aufgezeichneten Bewegungsabfolgen animiert. Während die statische Pose die Kommunikation gegenüber der realistischen Animation nur wenig verschlechterte, hatten aufgezeichnete Bewegungsabfolgen gravierend negative Auswirkung auf die Kommunikation der Teilnehmer. Dies erklärt sich durch die verwirrende Wahrnehmung, dass die nonverbale Kommunikation des Gegenübers nicht zur verbalen Kommunikation zu passen scheint. Hier zeigt sich wieder der in Abschnitt 3.3 angesprochene Uncanny-Valley Effekt. Während die realistische Gestik des Avatars der Kommunikation generell zuträglich ist, wirkt sich ein inkohärentes Verhalten des Avatars fatal auf die Akzeptanz des Menschen aus.

Guadagno et al. [30] haben untersucht wie sich die Abbildung von Mimik bei animierten Avataren auf die Kommunikationsqualität und das Gefühl der Präsenz auswirkt. Dazu wurde ein virtuelles Beratungsgespräch simuliert, in dem Testpersonen als virtueller Avatar durch

einen Avatar beraten werden. Der Berater wurde dabei durch einen Wissenschaftler über ein Motion-Capturing System animiert.

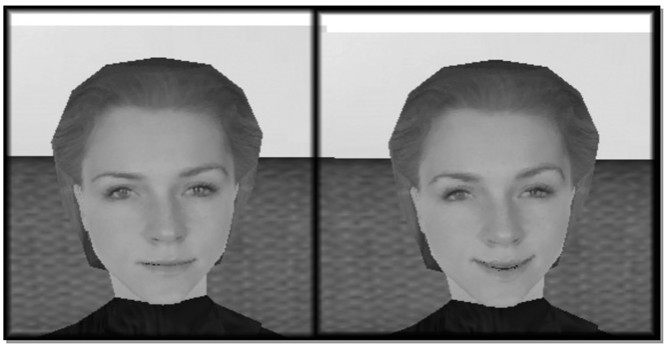


Abbildung 3.14: Berateravatar. Links neutral, rechts lächelnd. Quelle: [30].

Dieser konnte in passenden Situationen den Berater Avatar wie in Abbildung 3.14 dargestellt lächeln lassen. Diese Funktion wurde für den Berater unsichtbar aktiviert oder deaktiviert. So wusste der Berater nicht ob seine Lächelintention ausgeführt wird oder nicht. In einer anschließenden Befragung beurteilten die Testpersonen den Berater Avatar durchgehend positiver wenn die Lächeln Funktion aktiviert war. Die Arbeit schließt daraus, dass die durch das Lächeln des Avatars ausgelöste Empathie die Akzeptanz des virtuellen Gegenübers als Menschen steigert, und somit die abgefragten Empfindungen wie Vertrauen und Offenheit, Kopräsenz oder Behaglichkeit positiv beeinflussen.

3.4 Fazit

Die in Abschnitt 3.2 beschriebenen Arbeiten zeigen, dass Echtzeit-Rekonstruktion generell als machbar gilt. Die Herangehensweisen und die erzielten Ergebnisse unterscheiden sich jedoch abhängig vom Verwendungszweck relativ stark. Dies beginnt bei der Art der Sensorik welche Tiefenbild-Kameras beinhalten kann [20, 21, 22] oder aber lediglich auf vielen RGB- und iR-Kameras basiert [27]. Während das Erste ein günstige und einfache Möglichkeit darstellt, an dreidimensionale Daten zu gelangen, ermöglicht das Zweite durch die schiere Masse an Sensorik einen sehr detaillierten, fast realitätsgetreuen Avatar. Allerdings geht hierbei durch die benötigte Rechenkapazität sowie Berechnungszeit die Echtzeitfähigkeit verloren.

Ein weiterer signifikanter Unterschied ist die Art der erzeugten Datenstruktur. Während Polygongitter, welche durch die Computergrafik gut verstanden sind, eine nachvollziehbare Wahl sind [20, 21], werden zunehmend auch volumetrische Datenstrukturen wie Distanzfunktionen

verwendet [22]. Der Vorteil dieses Verfahrens ist die Möglichkeit die beschränkte Perspektive einer einzigen Kamera über die Zeit zu verändern, und so das gescannte Modell immer weiter zu vervollständigen und zu verfeinern und eine recht hohe geometrische Auflösung zu erreichen. Zunächst eignete sich das Verfahren aber nur bedingt für die Rekonstruktion von Menschen, die nur schwer bewegungslos sein können. Dieses Problem wurde kürzlich durch die Einführung eines Verformungstrackers eliminiert, welcher durch Rückrechnung der Verformung eine Integration in das statische Modell erlaubt und das statische Modell zur Anzeige wieder dynamisch verformen kann.

Aus der technischen Perspektive ist ein realistischerer Avatar gleichbedeutend mit einer höheren Rechenlast bei der Verarbeitung der Sensordaten. Zwar ist ein fast fotorealistischer Avatar mit heutigen Methoden durchaus erreichbar, jedoch erfordert die Methode einen enormen Rechenaufwand, welcher sich nicht für eine Echtzeit Anwendung eignet [27].

Auf der anderen Seite zeigen die in Abschnitt 3.3 vorgestellten Arbeiten die den Menschen in VR/AR untersuchen, dass die Kommunikation wesentlich natürlicher wird, je realistischer der Avatar darstellbar ist. Die untersuchten Einzelaspekte Blickrichtung [29], Gestik [9] und Mimik [30] wiesen bei realistischer Darstellung durchweg positive Effekte auf die Kommunikation zwischen Mensch und Avatar auf. Dies würde zunächst die Zielsetzung einer möglichst detailgenauen Abbildung eines Avatars in einem kollaborativem, tele-immersivem System bestätigen. Allerdings bestätigen die Arbeiten auch den Effekt des Uncanny-Valley wenn bei steigendem Avatarrealismus zum Beispiel das Verhalten unrealistisch wirkt [9, 29]. Ähnliche Effekte dürften auch bei kleinen Fehlern in einer nahezu perfekten Rekonstruktion auftreten.

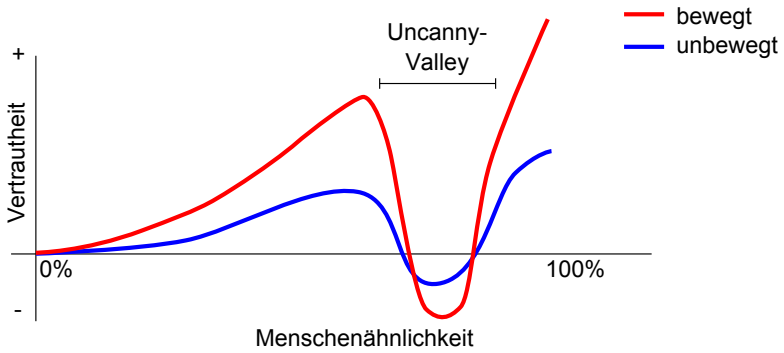


Abbildung 3.15: Uncanny Valley. Quelle: eigenes Werk, angelehnt an [31]

Hier wird deutlich, dass die Bestrebungen nach einer besseren Kommunikation durch möglichst realistische Darstellung eines Avatars relativ leicht durch Uncanny-Valley Effekte zunichte gemacht werden können. Der so anfänglich diametrale Widerspruch zwischen technischer Machbarkeit und gewünschter Qualität wird insofern etwas aufgebrochen als dass es unter

Umständen überhaupt nicht notwendig ist 100% Menschenähnlichkeit (siehe Abbildung 3.15) zu erreichen, sondern dass ein technischer machbarer Kompromiss unter Vermeidung von Uncanny-Valley Effekten unter Umständen sogar effektiver sein könnte.

4 Anforderungsanalyse

In dieser Arbeit sollen Avatare die mit verschiedenen Algorithmen zur Mesh Triangulation in verschiedenen Qualitäten erstellt werden gegeneinander in Bezug auf die Kommunikation mit einem Menschen verglichen werden. Da im Zentrum der Untersuchung die Frage nach einer sinnvollen Reduktion der 3D Modelldaten steht, spielt die verbale Kommunikation keine Rolle. Es wird also nur nonverbale Kommunikation des Avatars in Richtung eines Betrachters untersucht, um die Auswirkung verschiedener Rekonstruktionsmethoden zu bestimmen. Dazu sollen Testpersonen an einem durch [9] inspirierten Kommunikationspiel mit einem rekonstruierten Avatar teilnehmen. In diesem Spiel stellt der Avatar pantomimisch Begriffe dar, die von den Testpersonen erraten werden sollen. Da auch Methoden und Qualitäten untersucht werden sollen, die nicht in Echtzeit zu erreichen sind, ist es nicht möglich, diese Untersuchung in Echtzeit und interaktiv zu gestalten.

Um die Qualität von Avataren hinsichtlich der Kommunikation mit einem Menschen zu untersuchen wird ein System benötigt, welches Benutzern Sequenzen von 3D Modellen vorspielen kann, so dass diese die abgespielte Handlung interpretieren können. Es müssen vergleichbare Sequenzen von 3D Modellen in verschiedenen Qualitäten abgespielt werden können, damit die Rate der erfolgreichen Interpretation lediglich durch die Qualität der 3D Modelle und nicht zum Beispiel durch unterschiedliche Darstellung von Begriffen beeinflusst wird. Es sollen sowohl echtzeitfähige Rekonstruktionsmethoden zur Erstellung gröberer 3D Modelle verwendet werden als auch solche, die detailliertere 3D Modelle erstellen aber nicht mit interaktiven Frameraten funktionieren.

Die Anforderung der Vergleichbarkeit der zu erstellenden Sequenzen regt an, die verschiedenen Rekonstruktionen einer Sequenz auf identischen Sensordaten basieren zu lassen. Hieraus und aus der Anforderung auch nicht echtzeitfähige Rekonstruktionsmethoden zu verwenden folgt die Notwendigkeit, die Sensordaten zunächst aufzuzeichnen, um dann die Rekonstruktionen aus den aufgezeichneten Sensordaten zu vollziehen.

Zu den Bilddaten der Sensoren werden zusätzlich Kalibrierungsinformationen der Sensoren benötigt. Dazu zählen sowohl die räumliche Position und Orientierung der Sensoren um die Daten in ein globales Koordinatensystem zu integrieren, als auch die internen Linsenparameter

um die 2D Bilddaten in einen dreidimensionalen Raum zu projizieren. Während die internen Linsenparameter sich nicht verändern, muss die räumliche Position und Orientierung bei jeder Veränderung neu bestimmt werden. Da das Ausrichten von mehr als zwei Objekten mit sechs Freiheitsgraden im Raum für den Menschen keine triviale Aufgabe ist, wird ein System benötigt welches diese Aufgabe möglichst automatisiert.

Als Sensoren sollen für dieses System RGB-D Kameras verwendet werden um die Kalibrierung zu vereinfachen und um von der Vorverarbeitung der Tiefeninformationen zu profitieren.

5 Realisierung

Um die in Kapitel 4 definierten Anforderungen zu erreichen, wird für diese Untersuchung ein mehrstufiger Verarbeitungsprozess definiert. Zunächst werden die Rohdaten der zu bewertenden Szenen in Echtzeit aufgezeichnet. Die Rekonstruktion wird dann offline durchgeführt und die entstehenden Rekonstruktionsergebnisse zwischengespeichert. Zur Bewertung können die Rekonstruktionsergebnisse dann als Free-Viewpoint-Videos wiedergegeben werden.

5.1 Designentscheidungen

Das zu realisierende System wird auf Basis von openFrameworks [32] als Applikationsrahmen entwickelt. Für die Verarbeitung von Punktwolken wird die Pointcloud-Library [33, 34] verwendet. Diese stellt umfangreiche Werkzeuge zur Verarbeitung von Punktdaten bereit und bietet eine Anbindung von RGB-D Kameras über das OpenNI2 Framework [35].

5.2 Versuchsaufbau

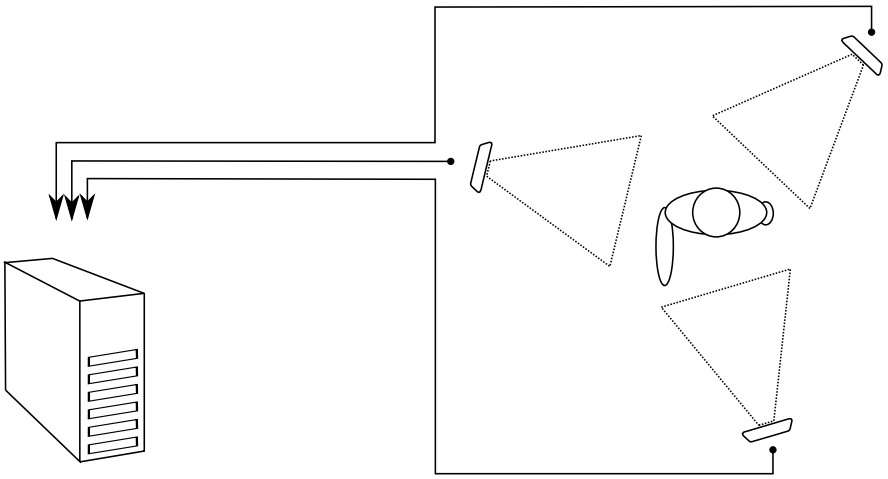


Abbildung 5.1: Aufbau des Scansetups. Quelle: eigenes Werk.

Zur Aufnahme der Sequenzen werden drei RGB-D Kameras wie in Abbildung 5.1 schematisch dargestellt radial in etwa gleichem Abstand von $1.8m$ um einen Mittelpunkt aufgestellt. Dabei sind die Kameras alle auf den zentralen Mittelpunkt ausgerichtet. Beim Aufstellen ist darauf zu achten, dass ein im Mittelpunkt stehender Mensch möglichst in allen Kameras voll abgebildet wird. Die Daten aller Kameras werden in einem Rechner zwischengespeichert.

5.3 Funktionsweise der Versuchssoftware

In diesem Abschnitt wird die Funktionsweise der Versuchssoftware beschrieben. Diese besteht aus mehreren separaten Programmen zur Kalibrierung, zur Aufnahme, zur Rekonstruktion und zur Wiedergabe der Rekonstruktionsergebnisse. Punkt-, Bild- und Kalibrationsdaten werden über Dateien ausgetauscht.

5.3.1 Kamerakalibrierung

Wie schon in Abschnitt 3.1 erwähnt haben RGB-D Kameras den Vorteil mit einem werksseitig vorkalibriertem und fest verbautem Farb- und Tiefenbildkamerapaar ausgeliefert zu werden. In einem Setup mit mehreren RGB-D Kameras müssen diese lediglich in ihrer räumlichen Position kalibriert werden.

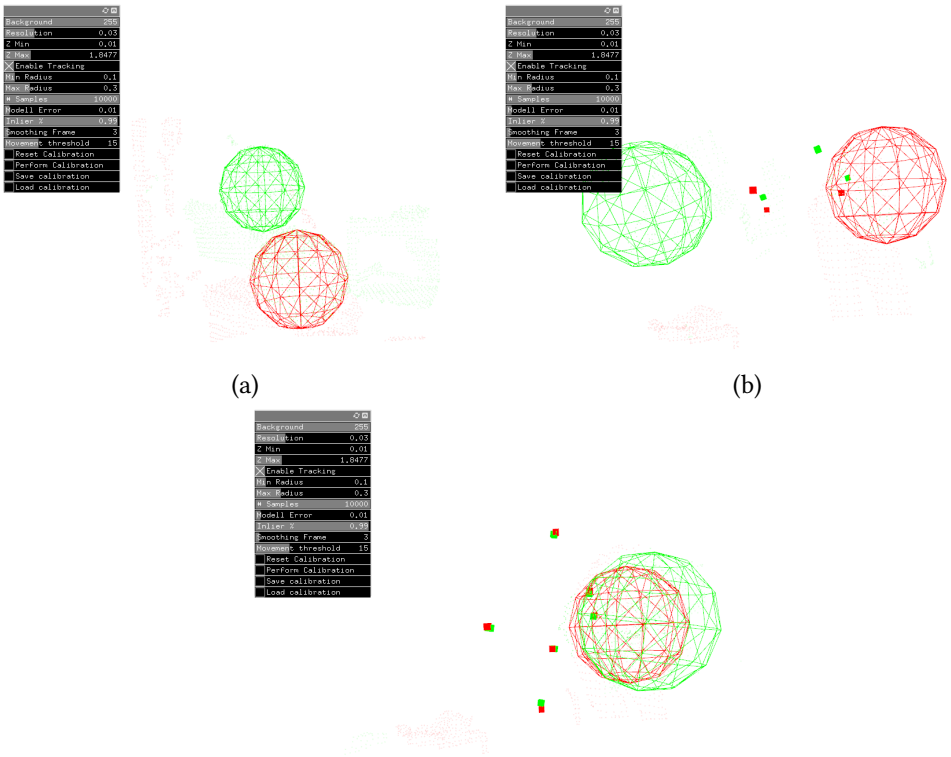


Abbildung 5.2: Kalibrierung mit zwei Kameras (Screenshots). Erkannte Kugeln (a), gespeicherte Kugelmittelpunkte (b), Punktwolken der Kugelmittelpunkte per ICP ausgerichtet (c). Quelle: eigenes Werk.

Dazu wurde in dieser Arbeit, durch [36, 37] inspiriert, eine Kalibrierungsmethode entwickelt, welche in allen Kameras simultan die Position einer Kugel bestimmt, und mit mindestens drei abgespeicherten Kugelpositionen die Transformation zwischen Kamerapaaren bestimmt. In der Punktwolke jeder einzelnen Kamera werden dazu mittels RANSAC diejenigen Punkte gefunden welche einem parametrischen Kugelmodel entsprechen. Dieses Kugelmodell wird mit einem Minimal- und einem Maximalradius initialisiert, und liefert nach dem Sampling die Parameter der Kugel mit der höchsten Trefferwahrscheinlichkeit (Abbildung 5.2a). Wird die Kugel für einen definierten Zeitraum nur innerhalb einer kleinen Distanz bewegt, wird der Kugelmittelpunkt in eine Punktwolke pro Kamera eingefügt (Abbildung 5.2b). Die Punktwolken der Kugelmittelpunkte unterscheiden sich dann abgesehen von Rauschen nur durch eine zu bestimmende Rotation und Translation. Mit einem kombiniertem Verfahren aus einer initialen Transformationsbestimmung mittels einer Sample-Consensus Variante und anschließender Verfeinerung durch einen ICP Algorithmus kann diese Transformation zwischen zwei Kameras

bestimmt werden. So werden paarweise die Transformation zwischen den Kameras bestimmt. Das Kalibrierungsprogramm führt diese Transformationsbestimmung nach jedem hinzugefügtem Messpunkt durch, sobald pro Kamera mindestens drei Messpunkte vorhanden sind. Der Erfolg der Kalibrierung kann durch den Benutzer durch Deckungsgleichheit der Messpunktwolken erkennen (Abbildung 5.2c). Die Kalibrierungsparameter werden in einer XML Datei gespeichert um von den Rekonstruktionsprogrammen verwendet werden zu können (siehe Abbildung 5.3).

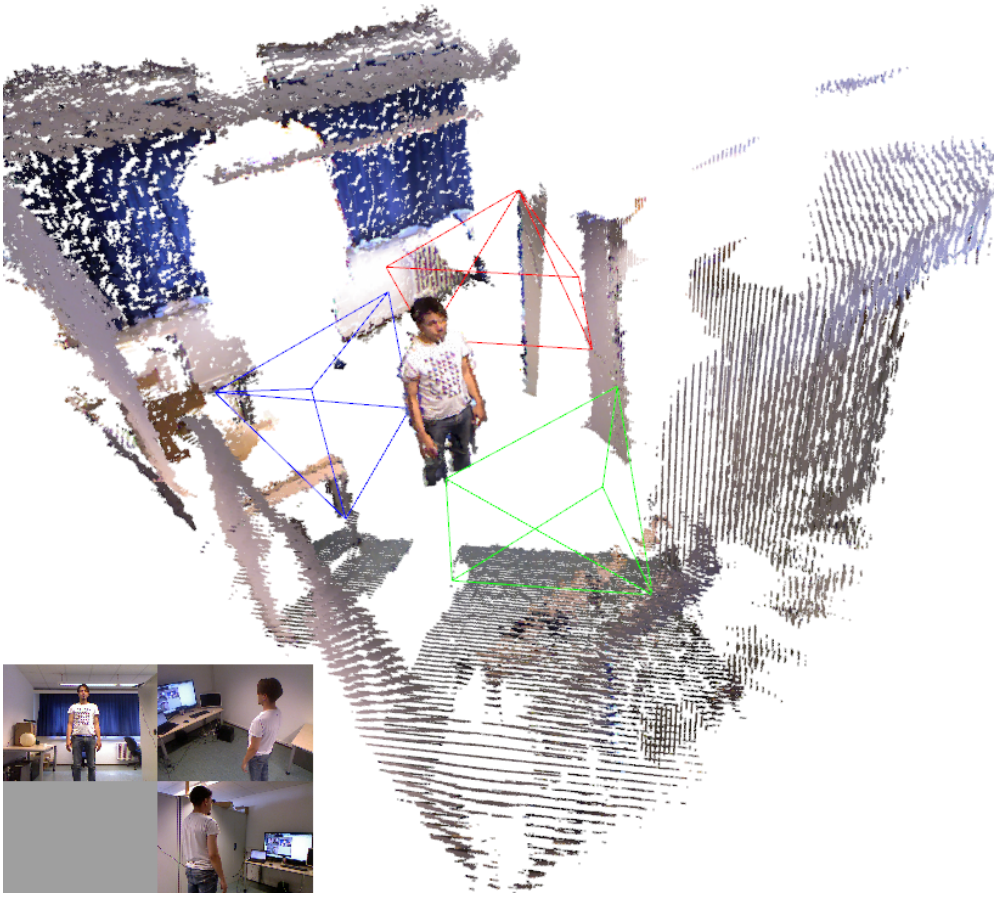


Abbildung 5.3: Mit Hilfe der Kalibrierungsparameter in globalem Koordinatensystem zusammengeführte Daten von drei Sensoren. Quelle: eigenes Werk.

5.3.2 Aufzeichnung

Um vergleichbare 3D Modelle mit unterschiedlichen Methoden zu erzeugen werden die Sensordaten zunächst aufgezeichnet. Für jede Kamera wird in jedem Frame die Punktwolke, sowie

das RGB Kamerabild gespeichert. Die Punktwolken werden dabei binär im PCD-Format der Pointcloud-Library gespeichert, die RGB-Bilder im PNG Format.

5.3.3 Verarbeitung

Die gespeicherten Punktwolken und Bilder werden dann in zwei verschiedenen Pipelines zu 3D Modellen verarbeitet. In der ersten, Echtzeit-fähigen Pipeline werden mit dem in Abschnitt 3.1 beschriebenen Verfahren pro Kamera ein Teil des 3D Modells erzeugt, und anschließend zusammengefügt. In der zweiten Pipeline werden zunächst die Punktwolken zusammengefügt und anschließend mit dem beschriebenen 3D-Triangulationsverfahren verarbeitet. Beide Pipelines sind für die Texturierung sowohl auf Vertex-Farben als auch auf Texturmapping eingerichtet. Im folgenden werden diese Pipelines näher beschrieben.

Organized Fast Mesh Pipeline

(a) (b) (c)

Abbildung 5.4: Organized Fast Mesh Rekonstruktionsergebnisse mit zwei Kameras. Hohe Punktdichte mit Vertexfarben (a), mittlere Punktdichte mit Vertexfarben (b) und mittlere Punktdichte mit Texturmapping (c). Quelle: eigenes Werk.

In dieser an [20] angelehnten Pipeline findet zunächst eine Verarbeitung pro Kamera statt. Zunächst wird durch Filterung der Punktwolke auf der Z-Achse der Hintergrund entfernt, so dass möglichst viele ungewünschte Artefakte verhindert werden. Hierbei wird darauf geachtet, dass die Punktwolke organisiert bleibt und so die Nachbarschaftsbeziehungen erhalten bleiben. Die verbliebenen Punkte werden nun mit Hilfe der Nachbarschaftsbeziehungen trianguliert. Dabei ist die Gitterweite konfigurierbar, so dass die geometrische Auflösung des Polygonmeshes eingestellt werden kann.

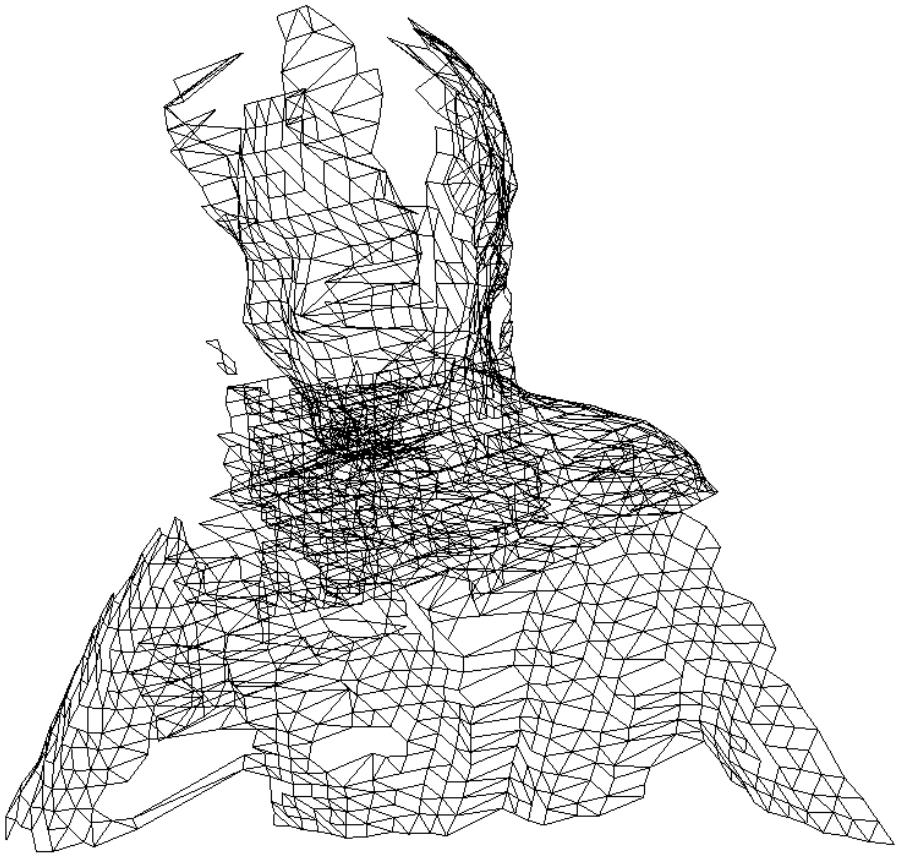
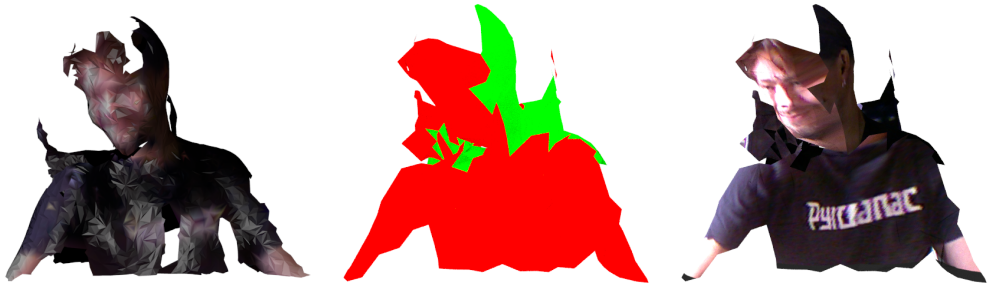


Abbildung 5.5: Wireframe des Organized Fast Mesh Rekonstruktionsergebnisses. Quelle: eigenes Werk.

Für die Punkte die nun zu dem Polygonmesh gehören, werden durch Projektion durch die Kameramatrix die Texturkoordinaten bestimmt. Hierbei muss die in Abschnitt 5.3.3 beschrieben Verarbeitung der Bilder beachtet werden. Die so entstehenden Teilmeshes für jede Kamera werden schließlich mit den durch die Kalibrierung erlangten Transformationen in einem gemeinsamen Raum zusammengebracht, und vereinigt. In dieser Arbeit wird im Gegensatz zu [20] auf die anschließende Polygonbereinigung verzichtet. In Abbildung 5.5 ist die charakteristische Struktur des so entstehenden Polygongitters dargestellt. Der Bereich in dem sich die rekonstruierten Halbschalen überschneiden ist durch die hohe Dichte an Kanten zu erkennen.

Greedy-Projection-Triangulation Pipeline



(a) (b) (c)

Abbildung 5.6: Greedy-Projection-Triangulation Rekonstruktionsergebnisse mit zwei Kameras. Mittlere Punktdichte mit Vertexfarben (a), Kamerazuordung der Polygone (b) und niedrige Punktdichte mit Texturmapping (c). Quelle: eigenes Werk.

In dieser Pipeline findet lediglich die Entfernung des Hintergrundes für jede Kamera einzeln statt, so lange die Punktwolken noch im Kameraraum sind. Danach werden die Punktwolken anhand der Kalibrierung in ein gemeinsames Koordinatensystem transformiert und in einer einzigen Punktwolke vereinigt. Nun wird mit einem Voxelgrid-Filter die geometrische Auflösung der Punktwolke festgelegt. Hierbei wird ein Voxelgitter von festgelegter Kantenlänge über die Punktwolke gelegt, und alle Punkte innerhalb eines Voxels auf einen Punkt im geometrischen Zentrum der Ursprungspunkte vereinigt. Anschließend werden für alle Punkte die Oberflächennormalen bestimmt und die Punktwolke mit einem Moving-Least-Square Verfahren polynomial geglättet (siehe [38]). Mit dem Greedy-Projection Algorithmus (siehe Kapitel 3.1) werden dann die Polygone trianguliert. Abbildung 5.7 zeigt die Struktur eines mit dieser Methode erstellten Polygongitters. Schließlich werden für die verbliebenen Punkte ebenso wie in der vorigen Pipeline Texturkoordinaten bestimmt. Dabei muss zunächst entschieden werden, welche Kamera die Textur für welches Polygon liefern soll. Dazu werden die Polygone anhand ihrer Sichtbarkeit und Verdeckung einer Kamera zugeordnet (siehe Abbildung 5.6b). Nicht zugeordnete Polygone werden mit Dummy Texturkoordinaten ausgestattet, die auf einen Texturpixel mit einer Alternativfarbe zeigen.

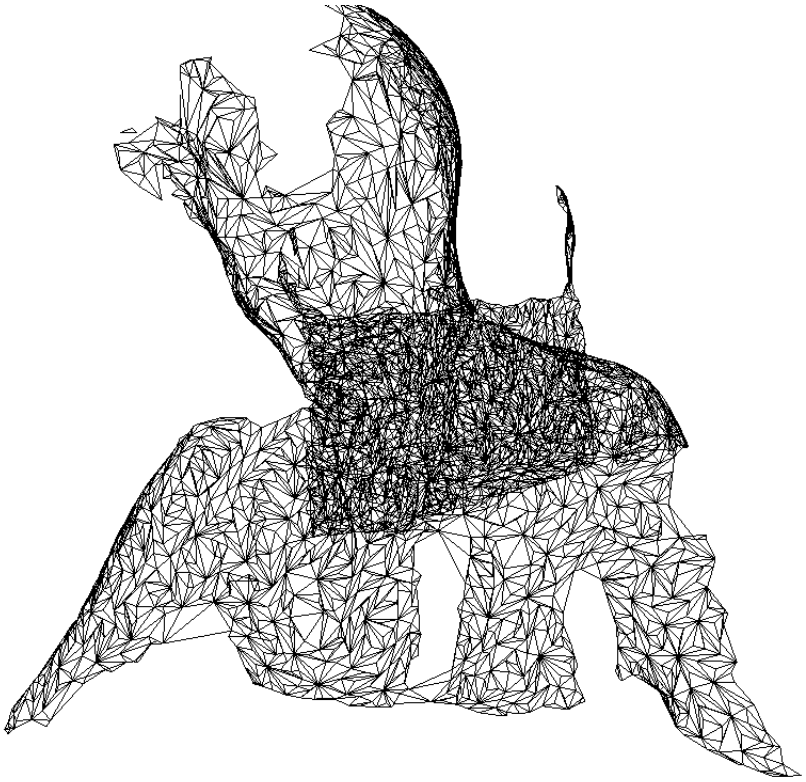


Abbildung 5.7: Wireframe des Greedy-Projection-Triangulation Rekonstruktionsergebnisses. Quelle: eigenes Werk.

Bildverarbeitung

Damit bei Texturen von mehreren Kameras eindeutige Texturkoordinaten vergeben werden können, werden die Texturen pro Frame zu einer einzigen zusammengefügt. Hierbei werden die Texturen in einer $2X2$ Matrix auf eine doppelt so große Textur gerendert und schließlich abgespeichert. Bei der Berechnung der Texturkoordinaten muss dann je nach Kamerazuordnung lediglich eine Verschiebung der Koordinate um eine Kamerabildbreite und / oder -höhe eingerechnet werden.

5.3.4 Wiedergabe

Für die Wiedergabe werden die gespeicherten Sequenzen zunächst geladen, um dann durch den Benutzer angesteuert in einer festen Framerate abgespielt zu werden. Dabei kann mit der Maus die Blickrichtung auf die Szene interaktiv verändert werden.

6 Untersuchung

Zur Untersuchung der Auswirkungen der verschiedenen Rekonstruktionsverfahren auf die nonverbale Kommunikation werden die Rekonstruktionsergebnisse einem Kommunikationsspiel mit Testpersonen unterzogen. Dieses Untersuchungsverfahren ist durch die Untersuchung von selbstanimierter Gestik durch Dodds et al. [9] inspiriert. Bei diesem Spiel geht es um das Erraten von Begriffen, welche pantomimisch durch den Avatar dargestellt werden. Die Testpersonen müssen die durch den Avatar dargestellte Handlung interpretieren und einen Begriff mit dieser Handlung assoziieren.

6.1 Versuchsbeschreibung

#	Begriff
0	Armdrücken
1	Haubentaucher
2	Luftballon
3	Raubüberfall
4	Umhang

Tabelle 6.1: Dargestellte Begriffe

Die Testpersonen versuchen, die fünf Begriffe in Tabelle 6.1 aus pantomimischen Beschreibungen zu erraten. Dabei wird einer der fünf Begriffe als Referenz durch den Autor in der selben Art und Weise dargestellt wie in den rekonstruierten Aufnahmen. Im Weiteren wird auf diese Teiluntersuchung als Mensch-Mensch Kommunikation verwiesen. Die restlichen vier Begriffe werden der Testperson als Avatar-Mensch Kommunikation über ein Free-Viewpoint Darstellungsprogramm vorgespielt. Dabei wird je eine hoch- und eine niedrig aufgelöste Rekonstruktion von beiden Pipelines wiedergegeben. Die Reihenfolge der Begriffe wird zufällig gemischt. Auch die Texturierung wird zufällig zwischen Vertexfarben und Texturmapping variiert. Die Testpersonen können die Sequenz selbständig abspielen, so oft sie wollen. Die Zeit

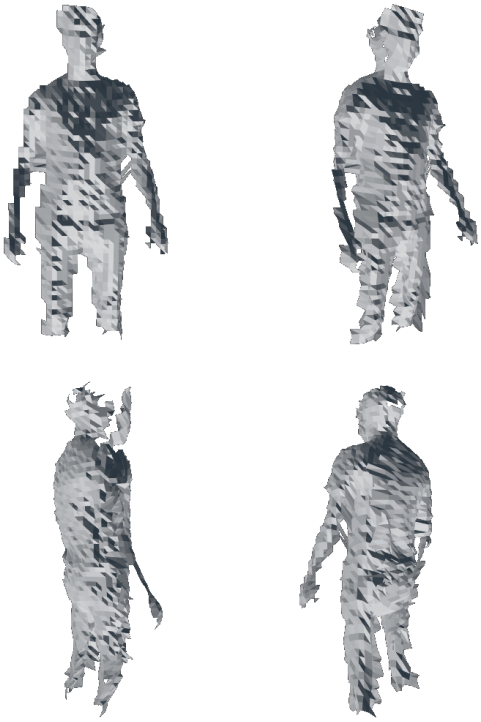
vom Beginn der Darstellung bis zum Aussprechen eines assoziierten Begriffes wird gemessen und notiert. Die Korrektheit einer Assoziation wird nicht auf den exakten dargestellten Begriff fixiert. Wenn der Teilnehmer die Assoziation anhand einer interpretierten Handlung erklären kann, die tatsächlich in die Art und Weise, wie der Begriff dargestellt wurde, herein interpretiert werden kann, wird die Assoziation als korrekt angesehen. Die verwendeten Versuche, sowie die benötigte Zeit werden zusammen mit dem Erkennungserfolg notiert. Für jeden der fünf Begriffe werden dazu im Vorfeld Sequenzen mit unterschiedlichen Methoden erstellt.

6.2 Erstellung der Test-Sequenzen

Die Test-Sequenzen wurden mit beiden Pipelines in jeweils zwei verschiedenen geometrischen Auflösungen erstellt. Die entstehenden 3D Modelle wurde inklusive Vertexfarben und Texturkoordinaten abgespeichert. Jeder Vertex beinhaltet also Orts- und Farbinformationen sowie Texturkoordinaten. Bei der Organized Fast Mesh Pipeline wird die geometrische Auflösung durch den Abstand der Pixel im Tiefenbild bestimmt welche als 3D Punkt in das Polygongitter aufgenommen werden. Bei der hochaufgelösten Variante wurde jeder Pixel des Tiefenbildes in das Gitter aufgenommen, bei der LQ Variante nur jeder Achte. Bei der Greedy-Projection Pipeline wird die Auflösung als räumlicher Abstand definiert. Dieser wurde in der HQ Variante auf einen Zentimeter und in der LQ Variante auf 3 Zentimeter gesetzt. In der Tabelle 6.2 sind die durchschnittlichen Zeiten für die Rekonstruktion pro Frame sowie die durchschnittliche Größe eines resultierenden Frames inklusive Textur mit einer durchschnittlichen Größe von $1400KB$ aufgelistet. Nachfolgend sind exemplarisch die Rekonstruktionsergebnisse eines identischen Frames mit der Organized Fast Mesh Pipeline (siehe Abbildung 6.1) sowie mit der Greedy-Projection Pipeline (siehe Abbildung 6.2) dargestellt.

Rekonstruktionsmethode	Frameberechnungszeit	Framegröße	# Vertizen	# Polygone
Organized Fast Mesh HQ	$1266ms$	$46,4MB$	$750K$	$250K$
Organized Fast Mesh LQ	$272ms$	$2300KB$	$15K$	$4K$
Greedy Projection HQ	$2064ms$	$7000KB$	$100K$	$30K$
Greedy Projection LQ	$557ms$	$1840KB$	$7K$	$2,5K$

Tabelle 6.2: Durchschnittliche Rekonstruktionsmetriken

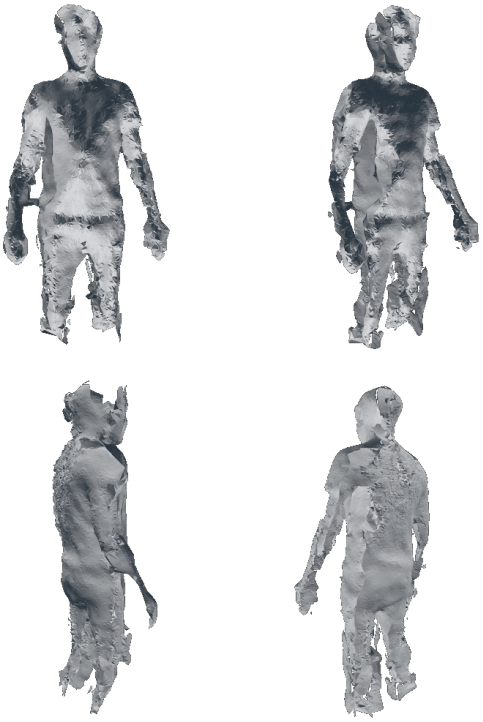


(a) Schrittweite bei Triangulierung: 8 Pixel im Tiefenbild

(b) Schrittweite bei Triangulierung: 1 Pixel im Tiefenbild

Abbildung 6.1: Organized Fast Mesh Ergebnisse ohne Texturierung. Quelle: eigenes Werk.

(a) Voxelgridauflösung: $3cm$



(b) Voxelgridauflösung: $1cm$

Abbildung 6.2: Greedy Projection Triangulation Ergebnisse ohne Texturierung. Quelle: eigenes Werk.

6.3 Auswertung

An der Auswertung nahmen 5 Frauen und 8 Männer im Alter von 26 bis 72 Jahren teil. Die Teilnehmer wurden über den Rahmen der Untersuchung und die Ziele der Untersuchung aufgeklärt. Dann wurde Ihnen der Ablauf des Experiments beschrieben und die Bedienung der Versuchssoftware gezeigt.

Anschließend versuchten die Teinehmer die in der Tabelle 6.1 gelisteten Begriffe anhand einer pantomimischen Darstellung zu erraten. Als erstes wurde ein Begriff durch Mensch-zu-Mensch Kommunikation dargestellt, anschließend wurden vier Begriffe in je einer der in Tabelle 6.2 gelisteten Rekonstruktionsmethoden abgespielt. Die Teilnehmer konnten sich die pantomimische Beschreibung beliebig oft ansehen und versuchten die dargestellte Handlung zu interpretieren und einen Begriff zu assoziieren. Der Erfolg einer Assoziation, die benötigte Zeit sowie die Anzahl der Abspielungen wurden zur Auswertung erfasst.

Die Erkennungsrate wird als Verhältnis der erfolgreich assoziierten Begriffe zu den gezeigten Begriffe berechnet. Für die Erkennungszeit sowie die Abspielwiederholungen werden nur die Messungen der erfolgreich assoziierten Begriffen berücksichtigt.

6.3.1 Vergleich der Avatar-Mensch Kommunikation gegen die Mensch-Mensch Kommunikation

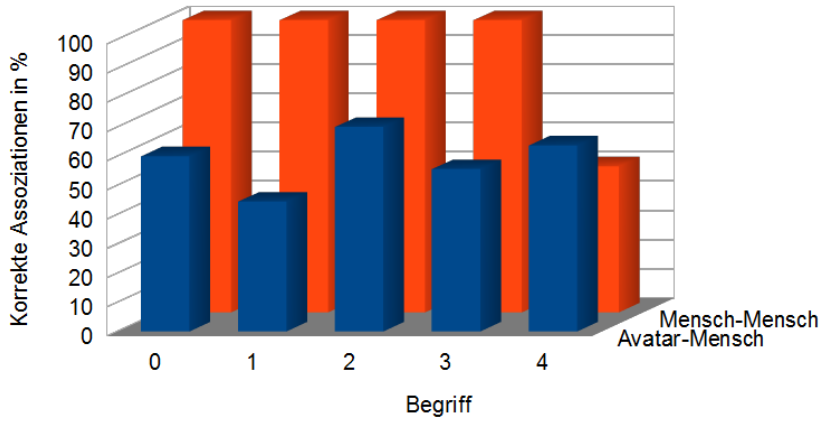


Abbildung 6.3: Erkennungsrate der Begriffe in der Avatar-Mensch und Mensch-Mensch Kommunikation. Quelle: eigenes Werk.

Durchschnittlich wurden Begriffe in der Mensch-Mensch Kommunikation zu $92,3\%$ und in der Avatar-Mensch Kommunikation zu $59,2\%$ mit einem vergleichbaren Begriff assoziiert. Für

die einzelnen Begriffe kann hier je Kommunikationsart eine vergleichbare Erkennungsrate festgestellt werden. Der Ausreißer in der Mensch-Mensch Kommunikation bei Begriff 4 erklärt sich durch die kleine Versuchsgruppe. Tatsächlich war dies der einzige Begriff, der in der Mensch-Mensch Kommunikation in einem Fall keine Assoziation hervorrief. Die Schwankungen in der Avatar-Mensch Kommunikation lassen hingegen auf die Schwierigkeit der Begriffe schließen. So wurde Begriff 1 von den Testpersonen häufig als schwer zu assoziieren genannt. Bei Begriff 3 wurde eine Bewegung einige Male fehlinterpretiert, was im Kontext zu einer höheren Schwierigkeit bei der Assoziierung eines Begriffes führte.

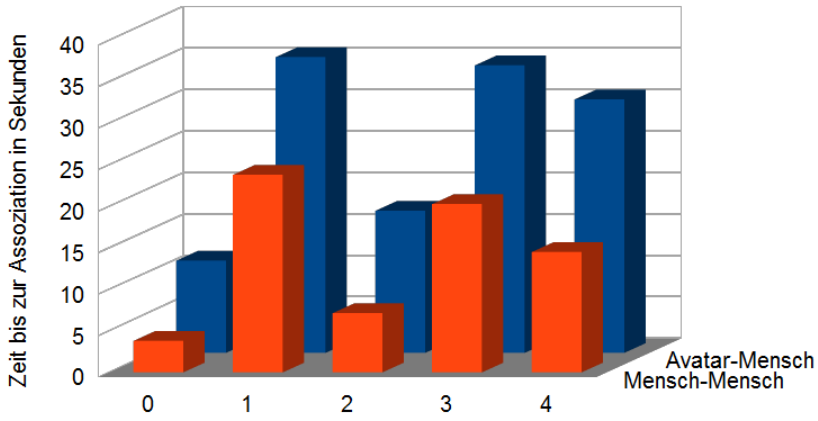


Abbildung 6.4: Erkennungszeit der Begriffe. Quelle: eigenes Werk.

Die Zeit bis zum Aussprechen eines Begriffes im Fall einer korrekten Assoziation scheint sich im kulminierten Vergleich zwischen Mensch-Mensch und Avatar-Mensch Kommunikation kaum zu unterscheiden. Durchschnittlich wurde in der Mensch-Mensch Kommunikation $14,86$ Sekunden benötigt, während in der Avatar-Mensch Kommunikation $16,16$ Sekunden benötigt wurde. Die Aufschlüsselung nach Begriff in Abbildung 6.4 zeigt zum einen für alle Begriffe eine vergleichbare Zeitdifferenz zwischen den beiden Kommunikationsformen und gibt zum anderen die oben angesprochene Schwierigkeit der Begriffe wieder. Mehrere Testpersonen merkten an, dass sie die meiste Zeit zum Finden eines Begriffes benötigten der zu der schon erkannten Handlung passte. Diese Beobachtung und eine starke Streuung der Zeiten weisen darauf hin, dass diese Metrik kein besonders gutes Bewertungskriterium ist. Der Vergleich der Rekonstruktionsmethoden in Abschnitt 6.3.2 zeichnet in diesem Aspekt ein etwas differenzierteres Bild.

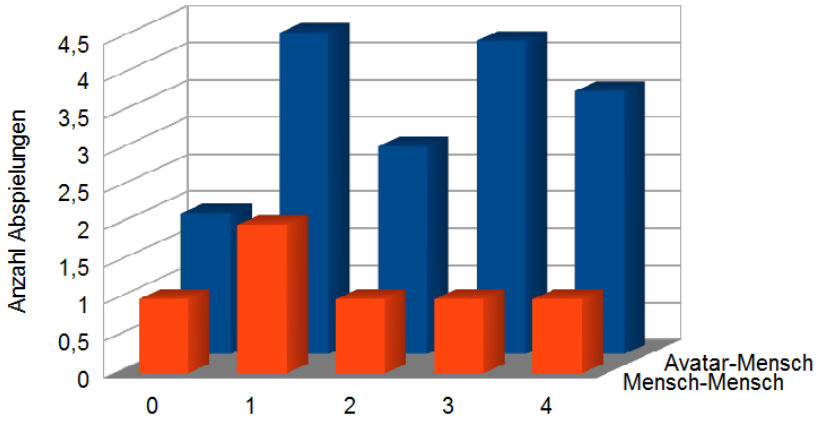


Abbildung 6.5: Anzahl Abspielungen nach Begriffen, Quelle: eigenes Werk.

Der Vergleich der Abspielwiederholungen zeigt, dass sich die Testpersonen in der Avatar-Mensch Kommunikation durchschnittlich mehr als doppelt so viele Wiederholungen ansahen wie in der Mensch-Mensch Kommunikation. Scheinbar waren sich die Testpersonen in der Mensch-Mensch Kommunikation durchschnittlich doppelt so schnell sicher, was sie gesehen hatten. Dies ist vermutlich auch durch die Tatsache zu erklären das die Untersuchung an einem konventionellen Monitor durchgeführt wurde. Die für den Menschen so fehlende stereoskopische Tiefenwahrnehmung wird durch nachträgliche Änderung der Perspektive ausgeglichen. Weiterhin reflektieren die Werte hier wieder die Schwierigkeit der Begriffe.

6.3.2 Vergleich der gewählten Rekonstruktionsqualitäten

In diesem Abschnitt werden die vier Rekonstruktionsvarianten aus Tabelle 6.2 hinsichtlich der Erkennungsrate der Begriffe, der dazu benötigten Zeit und der Anzahl der Abspielungen verglichen. Einige Testpersonen äußerten die subjektive Einschätzung, dass die Rekonstruktionsmethode wenig Einfluss auf ihre Fähigkeit den Begriff zu erraten hat. Dennoch zeigen die Daten ein auffälliges Muster.

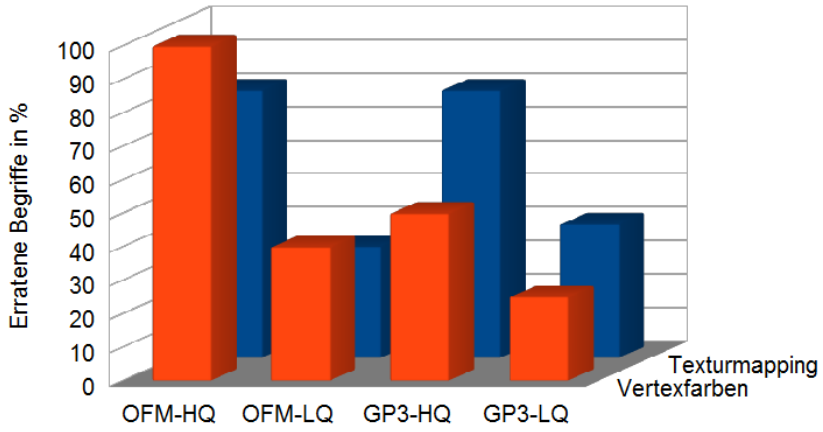


Abbildung 6.6: Vergleich der Erkennungsrate in Bezug auf die Rekonstruktionsmethode. Quelle: eigenes Werk.

Die Erkennungsrate bei den verwendeten Rekonstruktionsvarianten bilden den in Abschnitt 3.3 gefunden Konsens ab, dass die Kommunikation besser wird, je realistischer der Avatar dargestellt ist. Die detaillierteren Varianten beider Pipelines mit hoher geometrischer Auflösung (OFM-HQ und GP3-HQ) sind ihren jeweiligen niedriger aufgelösten Varianten überlegen. Eine Testperson erklärte, dass der Autor in den hoch aufgelösten Varianten sofort erkannt wurde, während in den niedrig aufgelösten Varianten nur noch die Handlung erkennbar war. Weiterhin wird klar, dass bei den Greedy-Projection Varianten die Texturierung mit Texturmapping einen wesentlichen Vorteil gegenüber der Texturierung mit Vertexfarben mit sich bringt. Dies ist durch die Zusammenfassung von Punkten im Voxelgridfilter bei der Auflösungsreduzierung zu erklären (siehe Abschnitt 5.3.3). Je gröber die Auflösung eingestellt wird, desto mehr Punkte werden inklusive ihrer Farben ausgemittelt, was die Textur zusätzlich zur Interpolation beim Rendern verwäscht. Ein weiterer Aspekt ist die etwas niedrigere Erkennungsrate der OFM-HQ Variante bei der Verwendung von Texturmapping im Vergleich zu der Verwendung von Vertexfarben. Erwartungsgemäß sollte der Unterschied hier marginal sein, da einzelne Modelle der beiden Varianten sich optisch kaum Unterscheiden sollten. Das Polygongitter der beiden Varianten ist identisch und die geometrische Auflösung ist so gering, dass beim Texturmapping fast jeder Punkt im Polygongitter auf einen eigenen Pixel im RGB Bild abgebildet wird. Es zeigt sich jedoch, dass die RGB Bilder mit einem leichten Delay ausgeliefert werden, so dass die Textur bei schnellen Bewegungen auf dem Polygongitter verschoben dargestellt wurde. Hier liegt die Vermutung nahe, dass dabei Uncanny-Valley Effekte die Ursache für die etwas

schlechtere Erkennungsrate sind. Dieser Fahler macht es schwer eine abschließende Beurteilung der Auswirkung der Texturierung durchzuführen.

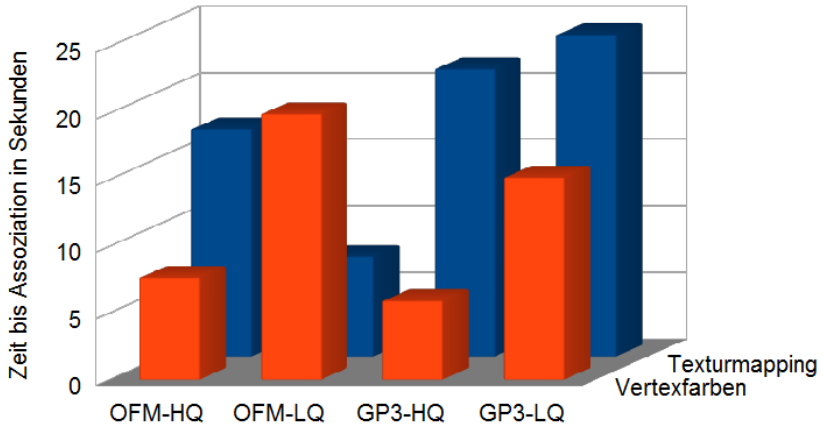


Abbildung 6.7: Vergleich der Erkennungszeit in Bezug auf die Rekonstruktionsmethode. Quelle: eigenes Werk.

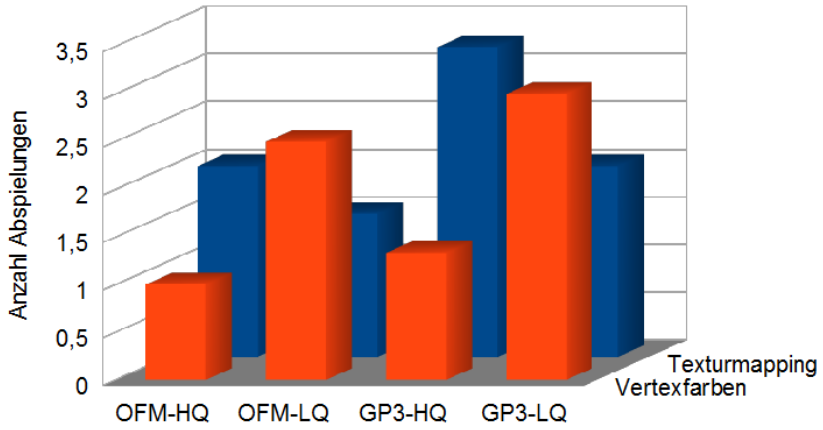


Abbildung 6.8: Vergleich der Anzahl von Abspielungen in Bezug auf die Rekonstruktionsmethode. Quelle: eigenes Werk.

Beim Vergleich der Anzahl von Abspielungen sowie der Zeit bis zur Assoziation fallen vor allem niedrige Werte bei den hochaufgelösten Varianten mit Vertexfarben sowie bei der niedrig aufgelösten Organized Fast Mesh Variante mit Texturmapping auf. Generell lassen diese

Werte aufgrund der in Abschnitt 6.3.1 angesprochenen Unsicherheiten nur wenige Schlüße zu. Allerdings kann ein gewisser Zusammenhang zwischen den beiden Metriken in Abbildung 6.7 und 6.8 festgestellt werden. Dabei fällt allerdings die GP3-LQ Variante auf. Während die Zeit zum Finden einer Assoziation unter der Verwendung von Texturmapping höher ist als unter der Verwendung von Vertexfarben, ist das Verhältnis bei der Anzahl der Abspielungen umgekehrt. Dies könnte durch die in Abbildung 6.2a sichtbaren Artefakte bei der Rekonstruktion der Gliedmaßen verursacht werden. Da sich die geometrisch grobe Struktur in jedem Frame verändert, entsteht eine Art für den Beobachter undefinierbare Form des Rauschens. Auch wiederholtes Abspielen bringt in dem Fall scheinbar keine Klarheit.

6.4 Fazit Auswertung

Die entwickelte Lösung konnte die in Kapitel 4 geforderten Anforderung weitestgehend erfüllen. Mit Hilfe des Kalibrierungsverfahrens können die extrinsischen Parameter mehrerer RGB-D Kameras innerhalb kurzer Zeit bestimmt werden. Während der Untersuchung reichte die Erfassung von maximal vier Kugelpositionen in allen Kameras aus um eine paarweise Transformationsbestimmung durchzuführen. Beim Aufnehmen der Sensordaten wird allerdings die erreichbare Framerate schon eingeschränkt, so dass die Sensordaten nur mit 7 FPS aufgenommen wurden. Hier muss für eine Echtzeitanwendung nachgebessert werden. Die Offline Rekonstruktion gestaltet sich als problemlos. Die Ergebnisse leiden vor allem unter der fehlenden Sensorsynchronisation und produzieren unter Randbedingungen geometrische Artefakte oder verrutschte Texturen. Durch die Entscheidung die verschiedenen Rekonstruktionsvarianten auf identischen Sensordaten basieren zu lassen, unterscheiden sich die Sequenzen lediglich durch die auf den Daten angewendeten Verfahren. Auch das Abspielen der rekonstruierten Sequenzen stellt sich als problemlos heraus. Die erfassten Metriken haben unterschiedlich starke Aussagekraft. Die Erkennungsrate der Begriffe scheint ein starker Indikator für die Kommunikationsfähigkeit des Avatars zu sein und bestätigt die in Abschnitt 3.3 beschriebenen Auswirkungen des Realismusgrades eines Avatars. Die starken Abweichungen in der Erkennungszeit sowie den Abspielungen deuten darauf hin, dass diese Werte nicht alleine durch die Kommunikationsfähigkeit des Avatars beeinflusst werden. Vielmehr spielt dort eine von Person zu Person sehr individuelle Fähigkeit der Assoziation offenbar eine größere Rolle. Hier bietet sich eine Modifizierung der Untersuchung an, die solche individuellen Faktoren ausschließen kann.

Trotz der hohen Erkennungsrate der mit der Organized Fast Mesh Pipeline in hoher geometrischer Auflösung rekonstruierten Sequenzen, ist diese Methode aufgrund einer Framegröße

von durchschnittlich $46,4MB$ eher schlecht geeignet für ein verteiltes Multi-User Szenario. Aus technischer Perspektive ist die Struktur des resultierenden Meshes relativ rau und abgestuft (siehe Abbildung 6.9). Durch die direkte Abbildung der Tiefenbildwerte auf 3D Punkte im Polygongitter sind die Quantisierungseffekte der Tiefenwerte deutlich sichtbar. Auch Rauscheffekte werden so direkt auf das Polygongitter abgebildet. An den Nahtstellen zwischen den Einzelgittern fällt die Überlappung zum Teil auf, da sich die erfassten Oberflächen durch Quantisierung und Rauschen nicht hundertprozentig gleichen.

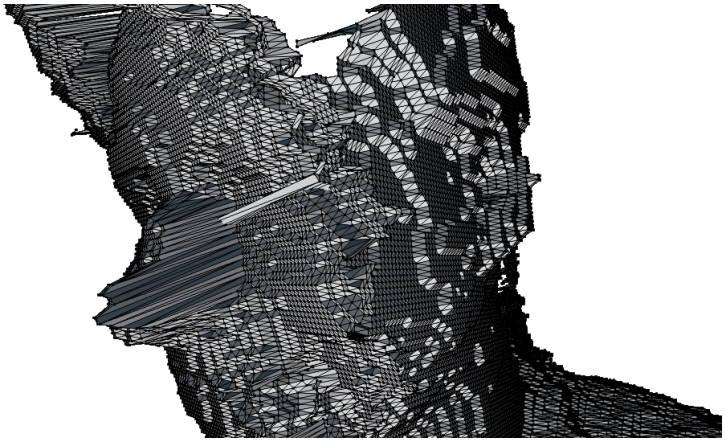


Abbildung 6.9: Struktur Organized Fast Mesh. Quelle: eigenes Werk.

Auch die geometrisch niedriger aufgelöste Variante scheint nach den Untersuchungsergebnissen keine Alternative zu sein. Die Erkennungsraten sind sowohl bei der Verwendung von Vertexfarben als auch mit Texturmapping deutlich niedriger als bei hoher Auflösung.

Die um den Faktor 10 kleinere Framegröße der Greedy-Projection Pipeline die bei hoher geometrischer Auflösung eine hohe Erkennungsrate ermöglichte, macht dieses Verfahren zunächst sehr attraktiv. Die Verarbeitungszeit von ca. 2 Sekunden pro Frame ist aber leider außerhalb des vertretbaren Rahmen für eine interaktive Anwendung. Hier besteht die Hoffnung, dass durch Weiterentwicklung der Hardware sowie Fortschritte in der Parallelisierung von Algorithmen die Rekonstruktionszeit weiter reduziert werden kann. Die Struktur des entstehenden Polygongitters ist wie in Abbildung 6.10 durch die Glättung zwar feiner Strukturen beraubt, weist aber eine wesentlich glattere Struktur auf. Wird die geometrische Auflösung bei diesem Verfahren allerdings zu groß, neigt es dazu zum Beispiel die Gliedmaßen nur als abstrakte Artefakte zu rekonstruieren.

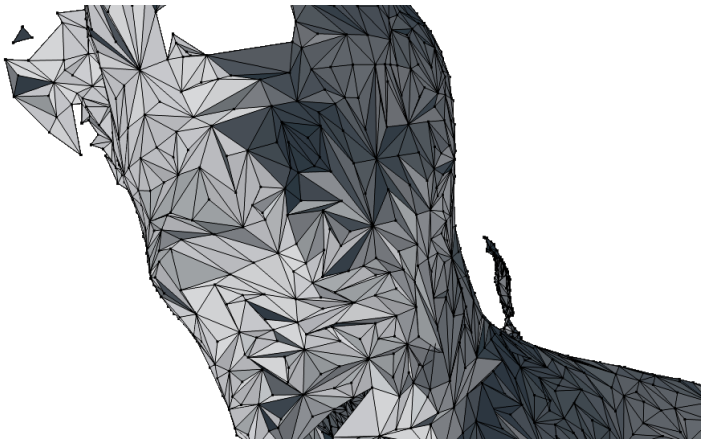


Abbildung 6.10: Struktur Greedy Projection Triangulation. Quelle: eigenes Werk.

Die Ergebnisse zeichnen ein differenziertes Bild der Auswirkungen von Triangulierungsverfahren, geometrischer Auflösung sowie Texturierungsverfahren. Die in Abschnitt 3.3 beschriebenen Thesen werden durch die Ergebnisse weitgehend bestätigt. Auch die vorliegende Untersuchung kann eine bessere Kommunikation durch einen realistischeren Avatar feststellen. Weiterhin hat sich auch gezeigt, dass kleine Imperfektionen von ansonsten hoch realistisch dargestellten Avataren starke negative Auswirkungen auf die Kommunikation haben können. Beispielsweise wird die Wirkung einer besseren Texturierung offenbar durch die leichte Asynchronität von Polygonmesh und Textur fast vollständig aufgehoben. Die Untersuchung zeigt, dass ausgerechnet die Verfahren mit den besten Erkennungsraten entweder zu große Frames produzieren, oder aber aufgrund der benötigten Berechnungszeit aktuell nicht in Echtzeit zu realisieren sind. Einen Echtzeitfähigen Kompromiss bieten die Organized Fast Mesh Variante mit Vertexfarben sowie die Greedy-Projection Variante mit Texturmapping in jeweils niedriger geometrischer Auflösung. Hierbei ist allerdings die Greedy-Projection Variante zu bevorzugen, da durch das Texturmapping auch Gesichtszüge erkennbar bleiben und somit die Mimik, welche in dieser Untersuchung eine untergeordnete Rolle spielt, als nonverbale Kommunikationsform zur Verfügung steht.

6.4.1 Probleme

Probleme bereitet zum Einen die bisher fehlende Synchronisation der Sensoren. Die nicht synchronisierte Auslösung der einzelnen Sensoren sorgt dafür, dass die Teilpunktwolken durch zeitliche Verzögerung nicht zusammenpassen. Bei schnellen Bewegungen ist aber auch die Verzögerung zwischen Punktwolke und RGB-Bild eines einzelnen Sensors schon problematisch,

da die Textur auf dem 3D Modell verschoben dargestellt wird. Dies zerstört jede noch so realistische Darstellung.

Ein weiteres Problem sind durch unvollständige Abdeckung der Sensoren verursachte Löcher im Polygongitter. Zwar könnte das Verfahren ohne weiteres um einige Kameras erweitert werden um die Abdeckung zu erhöhen, aber prinzipiell ist es immer möglich Teile der zu rekonstruierenden Oberfläche durch Verdeckung vor den Sensoren zu verstecken. Hier wäre es unter Umständen sinnvoll einen Bereinigungs- und Re-Triangulierungsschritt einzuführen. Hier könnte eine Untersuchung des in [20] beschriebenen Verfahren zu Bereinigung des Polygongitters Inspiration bieten.

7 Schluss

In dieser Arbeit wurde die qualitative Darstellung von nahe Echtzeit rekonstruierter dreidimensionaler Avatare im Hinblick auf ihre nonverbale Kommunikationsfähigkeit untersucht. Mit steigender Beschäftigung mit tele-immersiven, kollaborativen Anwendungen steigt der Bedarf nach kommunikationsfähigen Avataren, um die Kommunikation in virtuellen Welten möglichst natürlich zu gestalten.

In Kapitel 3 wurden zunächst einige Grundlagen der zur Echtzeitrekonstruktion verwendeten Technologien und Konzepte dargelegt. Eine Analyse verschiedener Ansätze zur Avatar Rekonstruktion zeigt, dass sowohl Echtzeitfähige als auch nahezu fotorealistische Rekonstruktion realisierbar ist. Jedoch schließen sich diese beiden Kriterien zur Zeit gegenseitig aus. Vorige Untersuchungen im Bereich der Wirkung verschiedener Faktoren von virtuellen Avataren auf die Kommunikation weisen auf den grundsätzlichen Vorteil höherer Realismusgrade hin. Allerdings zeigen sie auch den höheren Anspruch an das realistische Verhalten bei höherem Realismusgrad und die fatalen Auswirkung, wenn zwar die Darstellung realistisch ist, aber das Verhalten nicht.

In Kapitel 4 wurden die Anforderungen an eine Umgebung zum Untersuchen der Einflüße verschiedener Qualitätskriterien von Sequenzen rekonstruierter Avatare definiert. In Kapitel 5 wurde anschließend zunächst in Abschnitt 5.3 das entwickelte System beschrieben. Zur vereinfachten Kalibrierung wurde ein automatisches Kalibrierungsverfahren entwickelt, welches die extrinsischen Parameter aller beteiligten Kameras aus wenigen Positionen einer getrackten Kugel berechnet. Die Sensordaten werden durch eine Aufnahmekomponente aufgezeichnet, und können so in verschiedenen Variationen zu vollständigen 3D Modellen verarbeitet werden. Dabei wird sowohl die Texturierung durch Vertexfarben als auch durch Texturmapping möglich gemacht. Die rekonstruierten Sequenzen werden zur Untersuchung mit der Aufnahmeframerate wiedergegeben.

In Kapitel 6 wird dann die Untersuchung der rekonstruierten Sequenzen beschrieben. Zur Untersuchung der Avatar-Mensch Kommunikation wurde eine Kommunikationsspiel adaptiert, in welchem Testpersonen Begriffe anhand pantomimischer Beschreibung erraten sollten. Zur Bewertung wurde dabei die Rate der erfolgreichen Assoziation eines Begriffes, die benötigte

Zeit sowie die Anzahl der Abspielungen der jeweiligen Sequenz aufgezeichnet. Während die Erkennungsrate ein guter Indikator für eine erfolgreiche Kommunikation zu sein scheint, eigenen sich die Erkennungszeit sowie die Anzahl der Abspielungen aufgrund starker Streuung eher schlecht für eine Abschätzung. Zukünftige Untersuchungen sollten hier versuchen die Erkennung der Handlung von der Assoziation eines Begriffes zu differenzieren. Während beide Zeiten bei verschiedenen Personen sehr individuell sind, ist die Zeit, die für das Finden einer Assoziation gebraucht wird, kein Bestandteil der nonverbalen Kommunikation, sondern schließt eigentlich an die abgeschlossene Kommunikation an.

Die Auswertung der Untersuchung in Abschnitt 6.3 hat verschiedene Einflüsse der verwendeten Rekonstruktionsmethoden gezeigt. Erwartungsgemäß erreichen die Methoden mit dem höchsten Realismusgrad die besten Erkennungsraten. Diese sind aufgrund von Berechnungszeit oder entehender Datenmenge aktuell allerdings nicht in Echtzeit zu erreichen. Während die Untersuchung die Aussagen der in Abschnitt 3.3 untersuchten Arbeiten bestätigt, wurde eine Präferenz für des Greedy-Projection Verfahren unter Verwendung von Texturmapping herausgearbeitet. Dieses Verfahren produziert begünstigt durch die Oberflächenglättung ein wesentlich saubereres Polygongitter bei deutlich kleinerer Datenmenge als die Organized Fast Mesh Methode und profitiert besonders von der Texturierung durch Texturmapping. Hier sind allerdings noch Optimierungen nötig um die Eignung für ein Echtzeit-System zu verbessern.

7.1 Ausblick

Für die weitere Arbeit mit diesem Verfahren muss vor allem die Synchronisation der Sensordaten realisiert werden. Für eine frameweise Rekonstruktion ist es absolut notwendig, dass die zu einem Frame gehörenden Sensordaten zeitlich kohärent sind. Ein weiterer Aspekt der verbesserungswürdig ist, sind die durch fehlende Sensordaten entstehenden Löcher im Polygongitter. Hier würde sich möglicherweise eine Adaption des Verfahrens zur lokalen Re-Triangulierung von [20] anbieten. Die Parallelisierung von Teilaufgaben durch GPU Berechnung könnte die Performance der Rekonstruktionsalgorithmen noch etwas steigern. Mit diesen Verbesserungen würde es sich anbieten eine ähnliche Untersuchung mit mehr Testpersonen und einem breiterem Spektrum an Avatarqualitäten zu wiederholen.

In Zukunft wird es sich unter Umständen lohnen, von der frameweisen Rekonstruktion hin zu der Rekonstruktion durch temporale Integration zu schwenken. Die rasanten Entwicklung der Fusion Technologien [23, 22, 24] lassen diese Technologie auch für dynamische Echtzeit Anwendungen interessant werden. Vor allem durch die jüngste Entwicklung dynamische Szenen in ein statisches Modell zu integrieren, welches dann in Echtzeit verformt wird, lässt

sich die Idee einer tele-immersiven, kollaborativen Anwendung neu denken. So ist zum Beispiel denkbar, lediglich die Verformungsinformationen in Echtzeit zu erfassen und damit an einem anderen Standort eine vorher aufgezeichnetes hoch aufgelöstes Modell zu animieren.

Literaturverzeichnis

[1] Audi. The audi vr experience. Online [letzter Zugriff: 2016-07-24]. URL https://audi-illustrated.com/en/CES-2016/Audi-VR-experience.

[2] Volkswagen Group Research. Virtual technologies. Online [letzter Zugriff: 2016-07-24], . URL http://www.volkswagenag.com/content/vwcorp/content/en/innovation/Virtual_technologies.html.

[3] Volvo. The volvo xc90 experience is here. Online [letzter Zugriff: 2016-07-24]. URL http://www.volvocars.com/us/about/our-points-of-pride/google-cardboard.

[4] Seiko Epson Corp. Epson's moverio pro bt-2000 smart headset for professionals to experience a smarter workplace. Online [letzter Zugriff: 2016-07-24]. URL http://global.epson.com/newsroom/2015/news_20150623.html.

[5] IBM. Researchers from ibm, nokia and vtt bring avatars and people together for virtual meetings in physical spaces. Online [letzter Zugriff: 2016-07-24]. URL https://www-03.ibm.com/press/us/en/pressrelease/28644.wss.

[6] Microsoft Research. Holoportation. Online [letzter Zugriff: 2016-07-24], . URL https://www.microsoft.com/en-us/research/project/holoportation-3/.

[7] Enea Francesco Pavone, Gaetano Tieri, Giulia Rizza, Emmanuele Tidoni, Luigi Grisoni, und Salvatore Maria Aglioti. Embodying others in immersive virtual reality: Electro-cortical signatures of monitoring the errors in the actions of an avatar seen from a first-person perspective. *Journal of Neuroscience*, 36(2):268–279, 1 2016. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0494-15.2016.

[8] Ilias Bergström, Konstantina Kilteni, und Mel Slater. First-person perspective virtual body posture influences stress: a virtual reality body ownership study. *PloS one*, 11(2):e0148060, 2016.

[9] Trevor J. Dodds, Betty J. Mohler, und Heinrich H. Bülthoff. Talk to the virtual hands: Self-animated avatars improve communication in head-mounted display virtual environments. *PLoS One*, 6(10), 10 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0025759.

[10] François Giard und Matthieu J Guitton. Spiritus ex machina: Augmented reality, cyberghosts and externalised consciousness. *Computers in Human Behavior*, 55:614–615, 2016.

[11] Justine Cassell. Embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78, 2000.

[12] D. Holz und S. Behnke. Fast range image segmentation and smoothing using approximate surface reconstruction and region growing. In *Proceedings of the International Conference on Intelligent Autonomous Systems (IAS), Jeju Island, Korea*, 2012.

[13] Zoltan Csaba Marton, Radu Bogdan Rusu, und Michael Beetz. On Fast Surface Reconstruction Methods for Large and Noisy Datasets. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, May 12-17 2009.

[14] Oleg Alexandrov. A more complicated set (top) and its signed distance function (bottom, in red). Online [letzter Zugriff: 2016-07-24]. URL https://en.wikipedia.org/wiki/Signed_distance_function#/media/File:Signed_distance2.png.

[15] B. Curless und M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996.

[16] Hugues Hoppe, Tony DeRose, Tom Duchamp, John McDonald, und Werner Stuetzle. *Surface reconstruction from unorganized points*, volume 26. ACM, 1992.

[17] Ming Zeng, Fukai Zhao, Jiaxiang Zheng, und Xinguo Liu. A memory-efficient kinectfusion using octree. In *Proceedings of the First international conference on Computational Visual Media*, CVM'12, pages 234–241, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-34262-2. doi: 10.1007/978-3-642-34263-9_30. URL http://dx.doi.org/10.1007/978-3-642-34263-9_30.

[18] P. J. Besl und H. D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, Feb 1992. ISSN 0162-8828. doi: 10.1109/34.121791.

[19] Rufael Mekuria, Michele Sanna, Stefano Asioli, Ebroul Izquierdo, Dick C. A. Bulterman, und Pablo Cesar. A 3d tele-immersion system based on live captured mesh geometry. In *Proceedings of the 4th ACM Multimedia Systems Conference*, MMSys '13, pages 24–35, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1894-5. doi: 10.1145/2483977.2483980. URL http://doi.acm.org/10.1145/2483977.2483980.

[20] D.S. Alexiadis, D. Zarpalas, und P. Daras. Real-time, full 3-d reconstruction of moving foreground objects from multiple consumer depth cameras. *IEEE Transactions on Multimedia*, 15(2):339–358, 2013. ISSN 1520-9210. URL 10.1109/TMM.2012.2229264.

[21] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, und Andrew Wilson. Room2room: Enabling life-size telepresence in a projected augmented reality environment. ACM - Association for Computing Machinery, March 2016. URL http://research.microsoft.com/apps/pubs/default.aspx?id=262648.

[22] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, und Andrew Fitzgibbon. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 559–568, Santa Barbara, California, USA, 2011. ACM. ISBN 978-1-4503-0716-1. doi: 10.1145/2047196.2047270. URL http://doi.acm.org/10.1145/2047196.2047270.

[23] Richard A Newcombe, Dieter Fox, und Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 343–352, 2015.

[24] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, und Shahram Izadi. Fusion4d: Real-time performance capture of challenging scenes. SIGGRAPH, July 2016. URL http://research.microsoft.com/apps/pubs/default.aspx?id=264321.

[25] Sean Fanello, Christoph Rhemann, Vladimir Tankovich, Adarsh Kowdle, Sergio Orts Escolano, David Kim, und Shahram Izadi. Hyperdepth: Learning depth from structured light without matching. CVPR, June 2016. URL http://research.microsoft.com/apps/pubs/default.aspx?id=264202.

[26] Dorina Thanou, Philip A. Chou, und Pascal Frossard. Graph-based compression of dynamic 3d point cloud sequences. *IEEE Trans. Image Processing*, December 2016. URL http://research.microsoft.com/apps/pubs/default.aspx?id=260847.

[27] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, und Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69, 2015.

[28] Masahiro Mori. The uncanny valley. *Energy*, 7(4):33–35, 1970.

[29] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, und M. Angela Sasse. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 529–536, New York, NY, USA, 2003. ACM. ISBN 1-58113-630-7. doi: 10.1145/642611.642703. URL http://doi.acm.org/10.1145/642611.642703.

[30] Rosanna E. Guadagno, Kimberly R. Swinth, und Jim Blascovich. Social evaluations of embodied agents and avatars. *Computers in Human Behavior*, 27(6):2380 – 2385, 2011. ISSN 0747-5632. doi: http://dx.doi.org/10.1016/j.chb.2011.07.017. URL http://www.sciencedirect.com/science/article/pii/S0747563211001555.

[31] Tobias K. Illustration of the "uncanny valley" with german text. Online [letzter Zugriff: 2016-07-24]. URL https://commons.wikimedia.org/wiki/File:Mori_Uncanny_Valley_de.svg.

[32] openFrameworks. openframeworks is an open source c++ toolkit for creative coding. Online [letzter Zugriff: 2016-07-24]. URL https://openframeworks.cc.

[33] Radu Bogdan Rusu und Steve Cousins. 3D is here: Point cloud library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 2011.

[34] Pointcloud Library. Pointcloud library website. Online [letzter Zugriff: 2016-07-24]. URL http://www.pointclouds.org/.

[35] Occipital. Openni2. Online [letzter Zugriff: 2016-07-24]. URL http://structure.io/openni.

[36] Minghao Ruan und Daniel Huber. Calibration of 3d sensors using a spherical target. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 187–193. IEEE, 2014.

[37] Aaron N Staranowicz, Garrett R Brown, Fabio Morbidi, und Gian-Luca Mariottini. Practical and accurate calibration of rgb-d cameras using spheres. *Computer Vision and Image Understanding*, 137:102–114, 2015.

[38] M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, und C. T. Silva. Computing and rendering point set surfaces. *IEEE Transactions on Visualization and Computer Graphics*, 9(1):3–15, Jan 2003. ISSN 1077-2626. doi: 10.1109/TVCG.2003.1175093.

Abbildungsverzeichnis

Glossar

AR: Augmented Reality, Seite 1

Body-Ownership: Bewusstsein über den Besitz des eigenen Körpers. In der Psychologie ein Teil des Selbstbewusstseins, welcher essentiell für die Verbindung zwischen körperlichen Sinnen und Emotionen / Empfindungen sind. Bei Betroffenen verschiedener neurologischer Erkrankungen ist dieses Bewusstsein oft gestört, Seite 1

Extrinsische Kameraparameter: Definieren die Position und Orientierung einer Kamera in einem dreidimensionalen Raum., Seite 11

Free-Viewpoint-Videos: Zeit-Sequenzen von 3D Modellen, welche aus beliebiger Position betrachtet werden können. Auch spatio-temporale Modelle oder 4D Videos genannt., Seite 5

HMI: Human Machine Interface, Seite 1

Intrinsische Kameraparameter: Beschreiben die Linseneigenschaften einer Kamera in Form von Fokallänge, Bildsensorformat, Bildmittelpunkt und anderen., Seite 7

Photogrammetrie Eine Reihe an Mess- und Auswerteverfahren um aus Fotografien 3D Modell zu bestimmen, Seite 6

RANSAC: Random Sample Consensus. Eine iterative Methode zum wahrscheinlichkeitsbarierten Bestimmen von Parametern aus beobachteten Daten mit vielen Ausreißern, Seite 25

RGB-D Kamera: Eine Multikamera Einheit die je einen Stream RGB- und Tiefenbilder liefert. Auch RGB+Depth Kamera genannt., Seite 6

VR: Virtual Reality, Seite 1

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 3. August 2016 Iwer Petersen