Evaluierung einer Wahrnehmungsschwelle von Kameratrackingfehlern beim Compositing realer und virtueller S3D-Videos

Dipl.-Ing. Thomas Lagemann, Dipl.-Ing. (FH) Sara Kepplinger, Dipl.-Ing. Mara Seupel, B. Eng. Tobias Tittelbach,
Technische Universität Ilmenau, Ilmenau, Deutschland, *thomas.lagemann@tu-ilmenau.de*

Kurzfassung

Dieses Paper präsentiert eine Studie zur Untersuchung der Akzeptanz von Kameratrackingfehlern in stereoskopischen Videosequenzen für den Einsatz in der *Stereo-3D* (*S3D*) Fernsehproduktion. Die klassische Betrachtungsweise von Fehlertoleranzen beim Kameratracking ist rein technisch motiviert. Im Zentrum dieser Studie wird deshalb aus zuschauerzentrierter Sicht die Fragestellung untersucht, ab welcher Ausprägung Trackingfehler in den sechs Freiheitsgraden der extrinsischen Orientierung der Kamera wahrgenommen, beziehungsweise als störend eingeschätzt werden. Dies geschieht unter Anwendung einer Variation der *Methode der eben merklichen Unterschiede* (*Method of Limits*).
Die Auswertung zeigt, dass Unterschiede zwischen den einzelnen Parametern existieren und dass das subjektive Empfinden von Trackingfehlern weitaus toleranter ist als die harten rechnerischen Grenzen.

1. Einleitung

Seit dem erneuten Aufkommen von *Stereo-3D* Kinofilmproduktionen wird auch im virtuellen Fernsehstudio stereoskopischer Inhalt produziert.
Für die Qualität des *Compositings* (dem Übereinanderlegen mehrerer Bildebenen) realer und virtueller Videoinhalte ist eine zuverlässige Kamerapositionsbestimmung unerlässlich. Neben den intrinsischen Parametern werden dazu die sechs extrinsischen Freiheitsgrade für Translation (X, Y, Z) und Rotation (Tilt, Pan, Roll) der Kamera ermittelt. Auf Basis dieser Parameter erfolgt die perspektivisch korrekte Berechnung des computergenerierten Hintergrundes in Relation zur aufgenommenen realen Szene. Die Genauigkeit, mit der die extrinsischen Parameter erfasst werden, hängt stark vom zugrundeliegenden Prinzip des Trackingsystems, vom Kalibrierungsstatus der Kameras und gegebenenfalls vom Aufnahmeszenario ab.

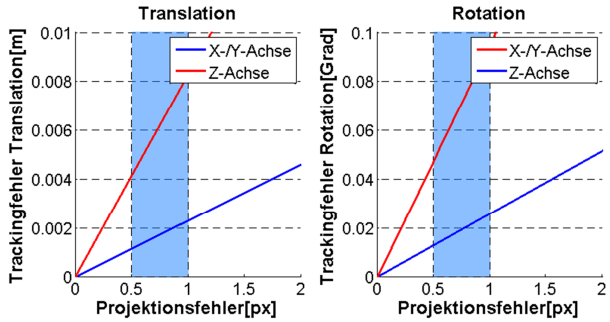

Bild 1 Zusammenhang zwischen Trackingabweichung und Projektionsfehler; für *Translation*(links) und *Rotation*(rechts)

Für die Produktion virtueller stereoskopischer Inhalte wird derzeit überwiegend auf Trackingsysteme zurückgegriffen, welche schon für die Produktion von monoskopischem SDTV entwickelt wurden.

Die Schwelle, ab der Trackingfehler sich im Bild auswirken, ist gemeinhin dann gegeben, wenn der geometrische Projektionsfehler zwischen 0,5 und 1 Pixel beträgt [1]. Ist also das Resultat des Trackingfehlers auf die Abbildung der virtuellen Szene größer als die Abmessung eines Sensorelementes, tritt ein Fehler auf. Für das in Bild1 gewählte Beispiel (2/3“ Sensor; Full HD; f = 12mm; Objektentfernung = 5m) liegt diese rechnerische Schwelle etwa zwischen 1 und 9 mm für die Translation und zwischen 0.1 und 0.9 Grad für die Rotation.
Die seit vielen Jahren im Produktionsumfeld genutzten Trackingsysteme überschreiten diese theoretischen Wahrnehmungsgrenzwerte jedoch teilweise um ein Vielfaches, ohne dass dadurch offensichtliche Probleme in der Akzeptanz des ausgestrahlten Inhalts beim Zuschauer registriert werden. In der Literatur angegebene physiologische visuelle Wahrnehmungsgrenzwerte decken sich mit oben genannter Grenze, ca. bei 0.5 - 1 Winkelminute [2].
Um sowohl bewährte als auch neu entwickelte Trackingsysteme gemäß ihrer Eignung zur Produktion von S3D-Inhalten zu beurteilen und konkrete Aussagen über die Wahrnehmbarkeit beziehungsweise den Störeinfluss von Trackingfehlern im *S3D*-Compositing machen zu können, untersucht die vorgestellte Studie wie Fehler im Kameratracking durch den Zuschauer wahrgenommen werden.
Die zentrale Fragestellung lautet dabei: Ab welcher Ausprägung werden Trackingfehler in den sechs Freiheitsgraden der extrinsischen Orientierung der Kamera wahrgenommen beziehungsweise als störend eingeschätzt.
Um Schwellenwerte definieren zu können sind *eben merkliche Unterschiede* interessant. Im Folgenden werden dazu drei Vorarbeiten vorgestellt, auf deren Basis derer die Anpassung bezüglich der oben genannter Fragestellung entwickelt wird. Nach der Vorstellung von

Test-szenario und -umgebung wird im Detail auf die Durchführung eingegangen. Im Anschluss an die Analyse der gewonnenen Daten werden die daraus resultierenden Ergebnisse vorgestellt und diskutiert.

1.1. Verwandte Arbeiten

Die *Methode der eben merklichen Unterschiede* wurde erstmals von G. T. Fechner vorgeschlagen [3] und auch bereits in ähnlichen Untersuchungen angewendet [4], [8]. Bei dem Verfahren geht es darum, eine Entdeckungsschwelle durch die allmähliche Erhöhung oder auch Verminderung der Intensität eines Reizes zu bestimmen um festzustellen, ab wann dieser gerade eben wahrnehmbar ist. Der Testteilnehmer reagiert darauf lediglich mit „Ja“ oder „Nein“ und deutet damit an, ob ein Unterschied festgestellt werden konnte. Dieses Verfahren wird mehrmals wiederholt um daraus den Mittelwert bilden zu können.
Die ITU [6] empfiehlt zwei Methoden um eine Wahrnehmungsschwelle festzustellen: a) die *Forced-choice Method* und b) die *Method of Adjustment*. Letztere ist aufgrund des technischen Aufwands der Testdatenerstellung (Rendering) und der damit einhergehenden nicht vorhandenen Echtzeitfähigkeit nicht anwendbar. Die *Forced-Choice-Methode* erschien für die vorliegende Fragestellung der betrachteten Anzahl an Parametern (dreimal Rotation, dreimal Translation) und der damit verbundenen notwendigen Versuchswiederholungen im Rahmen der vorgeschlagenen Testlänge von ca. 30 Minuten (siehe: [6]) nicht durchführbar.
Die *Methode der eben merklichen Unterschiede* hat den Vorteil, dass die Aufgabe für den Testteilnehmer recht einfach zu lösen ist.
Ähnlich zu McCarthy et al. [4] und Knoche et al. [8] wurde für die Feststellung einer Erkennungsschwelle die ursprüngliche Methode an die Fragestellung sowie an die technischen Gegebenheiten und Möglichkeiten angepasst. Die konkreten Änderungen werden im nächsten Abschnitt näher erläutert.

1.2. Anwendung der Methode

Im Unterschied zu den Untersuchungen von McCarthy [4] und Knoche [8] wird den Testteilnehmern bereits vorher mitgeteilt, ob die Qualität der gezeigten Sequenz besser oder schlechter wird. Zusätzlich ist bekannt, dass eine fehlerfreie Sequenz vorkommt. Damit entfällt die Notwendigkeit zu bewerten, ob die Qualität besser oder schlechter geworden ist. So müssen die Probanden bei einer Sequenz, in der die Qualität ansteigt nur angeben, ab wann sie eine Verbesserung wahrnehmen, indem sie zu diesem Zeitpunkt eine Taste drücken und umgekehrt. Falls die Testteilnehmer der Meinung sind, keine Veränderung festzustellen, können sie das mündlich mitteilen, ohne die Taste zu drücken. Bei der Untersuchung von Knoche [8] hingegen müssen die Probanden noch durch zwei Tasten unterscheiden, ob die Qualität besser oder schlechter wird.
Die Versuchsdurchführung ist in zwei Blöcke aufgeteilt, wobei die Teilnehmer jede Sequenz zweimal bewerten. Auf diesem Weg kann festgestellt werden, inwieweit sich die zwei Bewertungen derselben Sequenz unterscheiden.
Das Ziel ist es, die Werte aller Testteilnehmer zu vergleichen, um so für jede Sequenz einen Mittelwert bilden zu können, welcher dann die Entdeckungsschwelle darstellt.

2. Evaluation

Das Experiment besteht aus drei Teilen, wobei der mittlere Teil die eigentliche 30-minütige Testdurchführung ist. Das verwendete Testmaterial orientiert sich an der Fragestellung, der Kontrollierbarkeit und den technischen Erstellungsmöglichkeiten. Die Testumgebung orientiert sich an standardisierten Vorschlägen. Die Testteilnehmer wurden im universitären Umfeld rekrutiert.

2.1. Testmaterial

Das Testmaterial ist so gewählt, dass für jeden der 6 zu evaluierenden Parameter je zwei Videosequenzen gezeigt werden. Eine, in der sich der Fehler von Null bis zur Maximalschwelle verschlechtert und eine, in der eine Verbesserung bis zur Referenz (also Null) eintritt. Dazu kommt jeweils eine versteckte Referenz, die keinen Fehler enthält.
Um feststellen zu können, inwieweit sich zwei Bewertungen der gleichen Sequenz durch einen Probanden unterscheiden, wird jede Szene zweimal vorgespielt.
Die Länge der Testitems ist mit je 50 Sekunden definiert, um einerseits der Reaktionszeit der Testteilnehmer gerecht zu werden und andererseits nicht zu sprunghafte Änderungen des Fehlers zu erzeugen. Mit dieser Länge je Video ergeben sich somit 24 Minuten reines Videomaterial. Mit Pausen zwischen den einzelnen Sequenzen und Testabschnitten werden so die anvisierten 30 Minuten erreicht.
Um die volle Kontrolle über die Kameraparameter zu behalten und ein fehlerfreies Compositing als Referenz mit einbeziehen zu können, werden computergenerierte Vorder- und Hintergrundsequenzen verwendet. Auch wenn in diesem Fall beide Bilder künstlich erzeugt sind, so werden im Folgenden dafür die Begriffe reale und virtuelle Kamera verwendet. Dabei entspricht der Vordergrund der realen Kamera, die immer die korrekte Bewegung vorgibt, der Hintergrund der virtuellen Kamera, deren Bewegung auf den fehlerhaften Trackingdaten basiert.
Die Wahrnehmung von Trackingfehlern ist stark vom gezeigten Inhalt und damit von zahlreichen Faktoren wie Entfernung, Bildausschnitt oder Texturierung von

Vorder- und Hintergrund abhängig, die nicht alle in einer einzelnen Erhebung berücksichtigt werden können. Daher wurde eine typische Moderationsszene gewählt, in der ein direkter visueller Bezug zwischen realen und virtuellen Objekten erkennbar ist.
Inhalt der Szene (siehe Bild 2) ist die Totalansicht eines „Moderators“ mit Würfel in einem gleichmäßig texturierten Raum. Dadurch, dass die realen Videoobjekte direkten Kontakt zum virtuellen Hintergrund haben, sollen Diskrepanzen zwischen Vorder- und Hintergrund frühzeitig auffallen. Die ermittelten Grenzwerte sind demnach als kritische Mindestwerte anzusehen, die in anderen Szenarien gegebenenfalls größere Toleranzen erlauben.
Die Kamera vollzieht innerhalb der 50 Sekunden Sequenzdauer eine halbkreisförmige Bewegung von links nach rechts, der Fixpunkt bleibt dabei auf dem Moderator.
Für jeden Freiheitsgrad der Kamerabewegung wird nun eine kontinuierliche Variation des Fehlers von nicht vorhanden bis zur Maximalschwelle simuliert. Der Maximalwert wurde so gewählt, dass er eine extreme, für jeden sichtbare Ausprägung des Fehlers darstellt. Für die Translation beträgt er 30 cm, für die Rotation drei Grad.
Die reale und virtuelle Szene wird mit zwei verschiedenen Stereokameras berechnet, wobei die reale immer dem gleichen korrekten Animationspfad folgt. Der Fehler der virtuellen Kamera wird zur Position der realen hinzuaddiert. Die getrennt gerenderten Szenen werden anschließend zu einem Film zusammengefügt (siehe Bild 2). Die Auflösung der Videosequenzen beträgt 1920 x1080 bei 25 fps.

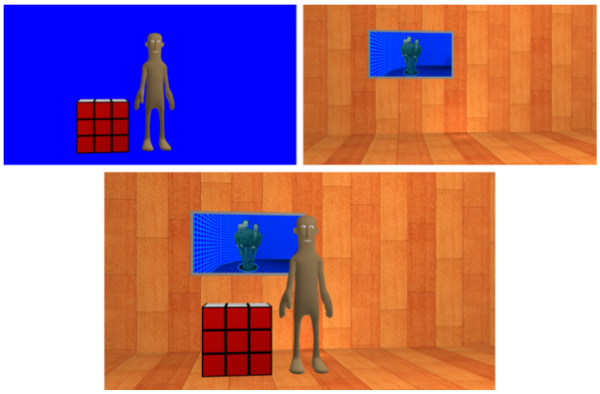

Bild 2 Zusammensetzung des Testmaterials: oben links: „Reale“ Kamera; oben rechts: „Virtuelle“ Kamera; unten: Compositing

2.2. Testumgebung

Die Laborumgebung, in der die Probandentests durchgeführt wurden, basiert auf den Standards ITU-R.BT.500 [6] und ITU-R.BT. 2160-2 [5]. Die Hintergrundbeleuchtung wurde unter Ausschluss von Tageslicht auf 200 Lux konstant gehalten. Der Betrachtungsabstand wurde gemäß *Preferred Viewing Distance (PVD)* bei gegebener Bildschirmhöhe auf rund 2m vorgegeben.
Die Zuspielung erfolgt mit der Software „Stereoscopic Player“ auf einem 24‘‘ Referenzmonitor mit Xpol-Zirkular-Polarisationssystem. Der Monitor ist über eine DVI-Schnittstelle mit dem Abspielrechner verbunden. Die Darstellung erfolgt mit voller HD-Auflösung (1920 x 1080, 25p) horizontal und halbierter Auflösung vertikal (line by line) für das linke und rechte Bild. Die Einstellungen des Monitors werden auf Werkseinstellung zurückgesetzt.

2.3. Testteilnehmer

Insgesamt nahmen 24 Personen (17 männlich, 7 weiblich) an dem Test teil. Bei den Teilnehmern handelt es sich um unbedarfte, in visueller Qualitätsbeurteilung nicht trainierte Personen, also keine Experten. Keiner der Teilnehmer ist farbenblind oder hat den Randot® SO-002 Stereotest (siehe auch 3.4.) nicht bestanden. Die Testteilnehmer sind zwischen 18 und 54 Jahre (Ø 27 Jahre) alt. Rund die Hälfte der Teilnehmer nutzen eine Sehhilfe (entweder Brille oder Kontaktlinsen).

2.4. Testdurchführung

Der Test teilt sich in 24 Einzeltests. Die Rahmenbedingungen sind bei allen Tests gleich. Dies beinhaltet die Art der Testdatenzuspielung und -präsentation sowie die Testumgebung und Beleuchtungssituation, welche von Standards (z.B. [5] und [6]) vorgegeben sind.
Die Testdauer betrug durchschnittlich 60 Minuten pro Teilnehmer. Zu Beginn des Tests wurde der Proband mit dem Testszenario und dem Testablauf vertraut gemacht. Neben dem Zweck der Untersuchung wurde auch die Dauer erläutert sowie auf Anonymität des Tests hingewiesen. Ein Pre-Test-Fragebogen diente vorab zur Erhebung der demografischen Daten und zur Feststellung des momentanen physischen Zustandes (Schwindel, Kopfschmerz, Übelkeit, ...) der Teilnehmer. Dafür wurde der Fragebogen zur 3D-Krankheit (Simulator Sickness Questionnaire (SSQ)) [12] vor und nach dem eigentlichen Testdurchlauf verwendet. Mit dem Randot® SO-002 Stereotest wurde die Fähigkeit des räumlichen Sehens untersucht und dokumentiert. Zum eigentlichen Test wurden ausschließlich Probanden mit intakter Binokularfunktion zugelassen. Im Test selber wurden die in Abschnitt 2.1 definierten Testvideos, bei denen sich die Qualität der Videos kontinuierlich verändert, in variierter und wiederkehrender Reihenfolge gezeigt, sowohl von gut nach schlecht als auch umgekehrt. Die Bewertung der Qualität war dem Testteilnehmer jederzeit möglich und erfolgte in dichotomer Weise anhand eines einfachen Eingabegerätes (in diesem Fall eine Computertastatur). Dabei wurde jeweils die Fragestellung beantwortet, immer dann zu bewerten (drücken), sobald sich etwas verbessert oder sobald sich etwas verschlechtert. Zusätzlich dazu wurden die Teilnehmer gebeten jeweils zu sagen, warum sie gerade eine Verschlechterung bzw. Verbesserung gewertet haben,

frei nach der *Methode des lauten Denkens (Thinking Aloud Method)* [9].
Die Reihenfolge wurde nach dem Lateinischen Quadrat (Latin Square Design) [10] (siehe dazu auch Vorschlag zu Randomisieren in der ITU-R BT.500 [6]) gewählt. Es wurde bewusst erwähnt, ob sich die Qualität der Testsequenz verschlechtert oder verbessert, um entsprechend nur Verbesserungen respektive Verschlechterungen durch Tastenklick festzuhalten. Nach der letzten Sequenz füllte der jeweilige Proband einen SSQ-Fragebogen aus und wurde abschließend zum Test befragt.

3. Analyse

Die Daten des SSQ vor dem Test werden mit den Ergebnissen nach dem Test verglichen. Die Daten zu den eben merklichen Unterschieden werden zuerst deskriptiv pro Kameraparameter analysiert. Hierbei wird festgestellt, ob es sich um normalverteilte Datensätze handelt und ob es Abweichungen oder Ausreißer gibt. Basierend auf der dabei festgestellten Art der vorhandenen Bewertungsdaten erfolgt eine Schwellenwertbestimmung, zum Beispiel durch Anwendung der *Varianzanalyse (ANOVA)* und des *Chi-Quadrat-Tests* für Zusammenhänge (siehe auch McCarthy et al. [4]) oder anderen geeignete nicht parametrische Analysemethoden (siehe auch Knoche et al. [9]). Diese Schwellenwerte ergeben sich aus der Bewertung pro gezeigtem Kameraparameter und dessen Qualitätsausprägung zu einer bestimmten Zeit. Die Werte werden zuerst in die Proportionen der akzeptablen Zeit nach McCarthy et al. [4] umgerechnet. Daraus wird die jeweilige Parametereinheit (bis zu 30 cm Abweichung für Translation und bis zu drei Grad Abweichung für Rotation) errechnet, welche die Abweichung von der rechnerisch korrekten Referenz beschreibt. Aufgrund dieser Informationen sowie der zusätzlichen deskriptiven Beschreibung der Qualität, die während der Wertung von den Teilnehmern mit erhoben wurde, erfolgt eine weitere Interpretation. In diesem Zusammenhang erfolgt ein Vergleich der Wertungen für den jeweiligen Kameraparameter gefolgt von einer Ergebnisdiskussion. Die Analyse geschieht mittels XlStat, einem MS Excel Add-In für statistische Zwecke.

4. Ergebnisse

Die Ergebnisse des SSQ erlauben es, alle gesammelten Datensätze mit zu berücksichtigen. Es haben zwar 11 der Teilnehmer (N=24) von überanstrengten Augen berichtet, allerdings konnte in der offenen Befragung danach festgestellt werden, dass die Ursache dafür eher in der Konzentration auf die Testsaufgabe als in einem 3D-Sehproblem zu finden ist. Dies wäre aber zusätzlich zu überprüfen.
Aus den Wertungen der Testteilnehmer lassen sich im Folgenden beschriebene Erkennungsschwellen feststellen. Lediglich die zweite Bewertungsrunde wurde hierbei und im Nachfolgenden berücksichtigt, da die erste Runde mehr Streuung aufweist und von einem Trainingseffekt ausgegangen werden kann. Bild 3 zeigt die zur Teilnehmeranzahl (N=24) relative allgemeine Wertungshäufigkeit in % des jeweils gezeigten Teststimulus (Translation: X-, Y-, und Z-Achse; Rotation: Pan (X), Tilt (Y), Roll (Z)), einmal für die zunehmende Abweichung und einmal für die abnehmende Abweichung des Trackingfehlers von der Referenz. Hierbei werden insgesamt alle abgegebenen Wertungen berücksichtigt.

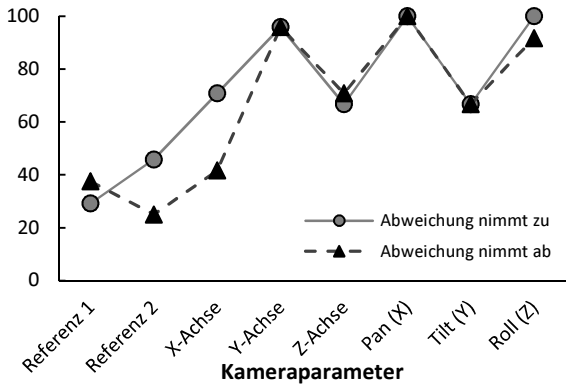

Bild 3 Relation der Wertungen pro gezeigtem Kameraparameter

Es ist zu erkennen, dass es unterschiedlich viele Wertungen für unterschiedliche Kameraparameter gibt. Die meisten Wertungen haben die Kameraparameter der Rotation Pan (X) sowie Roll (Z) mit jeweils 100 % für zunehmende Abweichung und 92 % für besser werdende Qualität, sowie die der Translation der Y-Achse mit je 96 % erhalten. Die schlechter werdenden Stimuli der X- und Z-Achse der Translation sowie Tilt (Y) der Rotation liegen jeweils um die 70 %, welches sich bei letzteren beiden bei Abnahme der Abweichung wiederholt. Der besser werdende Stimulus der Translation der X-Achse erhielt nur 42 % Wertungen. Die versteckten Referenzen haben relativ gesehen sehr wenige Wertungen erhalten. Dass hier überhaupt Wertungen abgegeben wurden, lässt sich mit der psychometrischen Funktion und dem Lapsus (falsche Wahl trotz richtiger Erkennung) [13] erklären. Eventuell spielt auch die soziale Erwartungshaltung, eine Wertung abgeben zu müssen, eine Rolle. Wenn man den Zeitpunkt der Abgabe der Wertung betrachtet, so liegt diese in allen Fällen bei den Referenzen zeitlich kurz vor Stimuli-Ende (siehe Bild 4 und Bild 5). Dies lässt ebenso auf die vorangegangene These schließen.
Die folgenden zwei Abbildungen zeigen die Grafiken der Mittelwerte der umgerechneten Akzeptanzzeiten. Einmal für die X-, Y-, und Z-Achse der Translation in Zentimeter (Bild 4) und einmal für die Kameraparameter der Rotation (X: Pan, Y: Tilt, Z: Roll) in Grad (Bild 5). Dafür wurde jeweils die erste Wertung pro Testteilnehmer, pro gezeigtem Stimulus berücksichtigt. Dieses erste Erkennen markiert den *eben merklicher Unter-*

schied sowie den Endpunkt der Akzeptanzzeit. Hierbei ist zu erkennen, dass es teilweise einen Unterschied in der Wertung von zunehmender zu abnehmender Abweichung gibt. Die Referenzen wurden jeweils ähnlich bewertet.

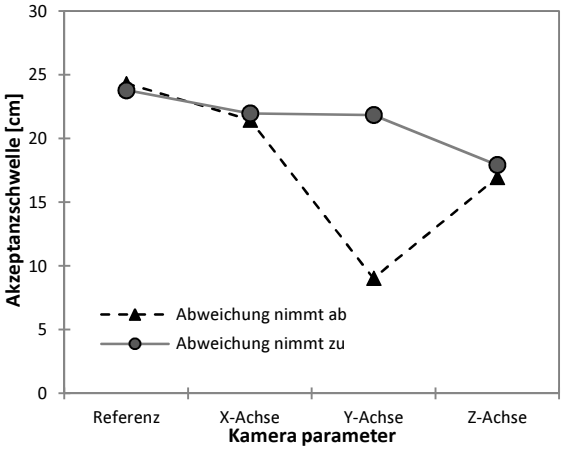

Bild 4 Wahrnehmung der Abweichung des gezeigten Kameratrackingfehler der Translation in cm

Die Abweichungen der Kameraparameter bei der Translation weisen einen signifikanten Unterschied zwischen zunehmender und abnehmender Abweichung bei der Y-Achse auf (Wilcoxon (T), α 0,05, p<0,0001). Eine Änderung wurde bei einer Verbesserung der Qualität der Y-Achse als früher erkennbar gewertet.

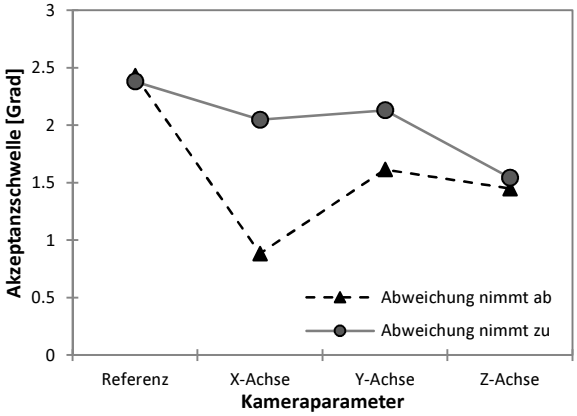

Bild 5 Wahrnehmung der Abweichung des Kameratrackingfehlers der Rotation in Grad

Die Abweichungen der Kameraparameter bei der Rotation zeigen vor allem bei Tilt(X) eine signifikant frühere (Wilcoxon (T), α 0,05, p<0,0001) und bei Pan (Y) und auch bei Roll (Z) eine leicht frühere Erkennung bei abnehmendem Fehler. Beide signifikanten Unterscheide(Trans (X), Tilt (X)) deuten auf eine hohe Empfindlichkeit gegenüber dem „Schwebeeffekt“ hin, der aus beiden Bewegungen resultiert. Generell ist zu bemerken, dass trotz relativ starker Abweichung von der rechnerischen Fehlerschwelle (siehe Kapitel 1) die akzeptable Zeit lang ist und der eben merkliche Unterschied bei einer hohen Abweichung auftritt. Unter Berücksichtigung der Anzahl der Wertungen (siehe Bild 3) lassen sich weitere Unterschiede feststellen: Fehler der Rotationsparameter Pan (X) und Roll (Z) und die Translation der Y-Achse werden früher und stärker wahrgenommen. Aus den qualitativen Daten, die in der abschließenden Fragerunde und während der Wertung mittels der Methode des lauten Denkens [7] erhoben wurden, lassen sich diese Ergebnisse bestätigen. Diese qualitätsbeschreibenden Aussagen wurden frei nach der Grounded Theory [11] sortiert und einer Clusteranalyse unterzogen. Das daraus entstandene Dendogramm (Bild 6) zeigt die daraus entstandene Klassenbildung nach der Häufigkeit der Nennungen. Das „Abschneiden“ (*Truncation*) bei einer Unähnlichkeit von etwa 30 führt zu zwei Gruppen, wobei die erste (links) homogener ist als die zweite (rechts). Dies wird bestätigt, wenn man die Intra-Klassen-Varianz betrachtet.

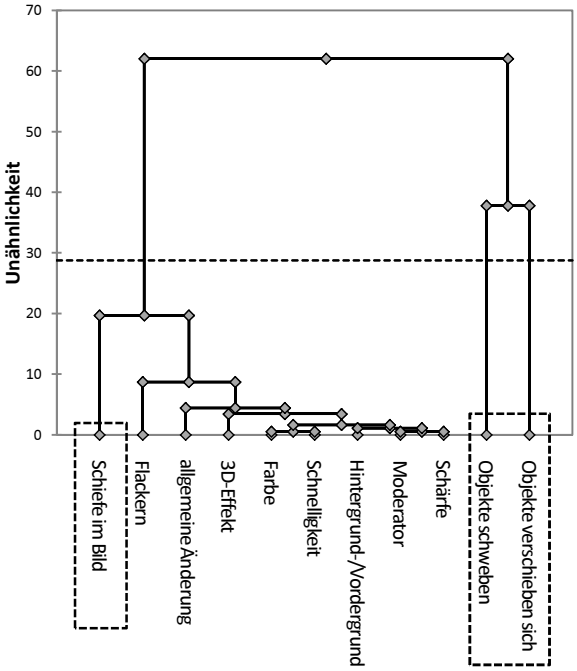

Bild 6 Dendrogramm aus der Clusteranalyse der beschreibenden Kommentare zur Qualitätsbeurteilung

Am häufigsten wurde das „Schweben“ des im Vordergrund dargestellten Moderators und Würfels, die „Verschiebung der Objekte“, sowie der „schiefe Raum bzw. Konfusionen der Hintergrund-/Vordergrund-Zusammensetzung“ beschrieben. Ebenso wurde von Unruhe, Flimmern und Flackern im Hintergrund berichtet. Diese Wertungen wurden für die Analyse und weitere Interpretation entfernt, da der Fokus auf den Kameraparameterabweichungen liegt. Der Grund dafür liegt in der Feststellung von Aliasingartefakten im Hintergrundbild, welche unabhängig von den sechs Kameraparameteränderungen auftreten. Zu erkennen ist ebenso eine Nennung weniger sogenannter 2D-relevanter Fehlerbeschreibungen wie Farbe, Schärfe und Schnelligkeit, obwohl sich diese Faktoren nicht änderten. Die-

se wurden in der Schwellenwertbestimmung ebenso wenig berücksichtigt. Ebenso lässt sich aus den qualitativen Aussagen schlussfolgern, dass ein Unterschied in der Qualität eher merklich ist, wenn sich die schlechte Qualität verbessert (Abweichung nimmt ab), aber leichter zu werten ist wenn sie sich verschlechtert, welches durch die Anzahl der Wertungen (Bild 3) und den Zeitpunkt der Wertungen (Bild 4 u. 5) bestätigt wird.

5. Zusammenfassung

Mit der vorgestellten Evaluationsmethode konnte eine erste Tendenz in der Wahrnehmung von Trackingfehlern ermittelt werden. Zu sehen ist, dass einzelne Abweichungen der drei Freiheitsgrade Translation (Y), Pan (X) und Roll (Z) der Rotation früher bemerkt und dann auch stärker wahrgenommen werden. Die niedrigen Wertungen für abnehmenden Fehler bei Pan(X) und Translation(Y) lassen auf eine größere Empfindlichkeit bezüglich der Bodenhaftung von Objekten schließen. Die Integration von qualitätsbeschreibenden Aussagen hat sich im vorliegenden Fall als sehr sinnvoll erwiesen.
Im Rahmen dieses Experiments konnten zusätzliche beeinflussende Indikatoren festgestellt werden, zum Beispiel ob eine Qualitäts-verbesserung oder -verschlechterung bewertet wird, und ob die im Fokus stehenden Kameraparameterfehler oder andere bewertet werden.
Im Vergleich zu den Wertebereichen des rechnerisch akzeptablen Trackingfehlers hat sich insgesamt gezeigt, dass Abweichungen erst sehr viel später wahrgenommen werden. Demnach sind, hinsichtlich der Bewertung der Qualität von Trackingsystemen, weitere Untersuchungen der zentralen Fragestellung, ab welcher Ausprägung Trackingfehler in den sechs Freiheitsgraden der extrinsischen Orientierung der Kamera wahrgenommen beziehungsweise als störend eingeschätzt werden, lohnenswert.

6. Ausblick

Das verwendete Testdesign hat gezeigt, dass der Einfluss von Gewöhnung und Erwartungshaltung der Testpersonen gegenüber den Testdaten neben der Verminderung durch versteckte Referenzen und randomisierter Präsentation noch genauer definiert und mitberücksichtigt werden muss. Außerdem wurde ein starker Trainingseffekt durch die erste Bewertungsrunde festgestellt, woraufhin diese für die kritische Analyse ausgeschlossen wurde. Das Einbeziehen einer expliziten Referenz kombiniert mit vergleichendem Testdesign kann dazu beitragen dem Probanden ein besseres Gefühl für den gesuchten Fehlereinfluss zu vermitteln.
Da subjektiv empfundene Videoqualität stark abhängig vom gezeigten Inhalt ist, sollten in nachfolgenden Tests vor allem verschiedene inhaltliche Szenarien verglichen werden. Dabei sollte Wert auf die Wahl von Szenarien gelegt werden, die wenige 2D-relevante Fehlereindrücke hervorrufen. In der Realität werden Trackingfehler nicht nur einzelne Freiheitsgrade beeinflussen. Somit sind komplexere, dynamisch variierende, Parameterkombinationen zu untersuchen.
Im Anschluss an die detaillierte Analyse der Fehlerwahrnehmung selbst bietet es sich an, Methoden zur Fehlerverschleierung zu untersuchen.

Literatur

[1] Moshkovitz, M. *The Virtual Studio. Focal Press.* 2000

[2] Mallot, H.A. *Sehen und die Verarbeitung visueller Informationen. Vieweg Verlagsgesellschaft.* Auflage: 2., überarb. und erw. Aufl. 2000 (25. Februar 2000).

[3] Fechner, G.T., *Elemente der Psychophysik Teil 1*, Breitkopf und Härtel, 1860.

[4] McCarthy, J., Sasse, M. A., & Miras, D. *Sharp or smooth? Comparing the effects of quantization vs. frame rate for streamed video.* In: Proc. CHI, 2004, S. 535-542.

[5] ITU Recommendation ITU-R BT.2160-2. *Features of three-dimensional television systems for broadcasting.* International Telecommunication Union, Genf, 2012.

[6] ITU Recommendation ITU-R BT.500-13. *Methodology for the subjective assessment of the quality of television pictures*. International Telecommunication Union, Genf, 2012.

[7] Nielsen, J. *Usability Engineering.* Academic Press, 1993. S.195.

[8] Knoche, H. McCarthy, J. Sasse, M.A. *How low can you go? The effect of low resolutions on shot types in mobile TV.* In: Multimedia Tools and Applications, Vol. 36, Issue 1-2, 2008, S. 145-166.

[9] Häder, M. *Empirische Sozialforschung: Eine Einführung.* VS Verlag für Sozialwissenschaften; Auflage: 2006 (26. Oktober 2006).

[10] Hinkelmann, K. Kempthorne, O. *Design and Analysis of Experiments*. Volume 1, Introduction to Experimental Design. 2nd Edition. John Wiley & Sons, 2008.

[11] Glaser, B, G. and Strauss, A., L. *The discovery of grounded theory: strategies for qualitative research.* Aldine, Chicago, USA, 1967.

[12] Kennedy, R. S. Lane, N. E. Berbaum, K. S. and Lilienthal ,M. G. *Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness,* The International Journal of Aviation Psychology, vol. 3, no. 3, pp. 203–220, 1993.

[13] Baird, J.C., *Sensation and judgment: complementary theory of psychophysics.* 1997, Mahwah, NJ: Lawrence Erlbaum Associates. 347.