

Evaluierung medizinischer Volumenrendering-Algorithmen durch empirische Studien

Wie gut ist ein Rendering-Algorithmus, der aus Daten Bilder generiert, die wiederum als Grundlage für eine medizinische Diagnose verwendet werden? In diesem Beitrag werden einerseits beispielhaft empirische Studien und ihre Ergebnisse vorgestellt, die solche Fragestellungen beantworten. Andererseits widmen wir uns auch der generellen Frage, welche Herausforderungen gemeistert werden müssen, um empirische Studien aufzustellen, die in Bezug auf Korrektheit medizinischer Diagnosen und ihre zugrunde liegenden Rendering-Algorithmen sinnvoll sind.

1. Motivation

Algorithmen der Computergrafik beschreiben vereinfacht die komplexe Interaktion von Licht und dreidimensionalen Objekten bzw. ihren Modellen; somit sind diese Algorithmen nur Approximationen der physikalischen Wirklichkeit. Approximationen in Rendering Verfahren werden zwar vom Menschen oft erkannt, aber nicht immer als problematisch wahrgenommen. Man erinnere sich an den Film *Toy Story* aus dem Jahre 1995, der als erster vollkommen durch Rendering-Algorithmen entstandener Film in Kinofilmlänge heute in seinen damals „realistischen“ Eigenschaften im Vergleich zu den mit biegsameren Objekten ausgestatteten Nachfolgefilmen sehr steif wirkt. Ähnlich vergleicht sich ein Computerspiel aus dem Jahre 2012 mit einer älteren Vorgängerversion ohne Schattenwurf oder in wesentlich niedrigerer Auflösung. Das heißt, wir entschuldigen grobe Approximationen so lange, bis wir etwas Besseres sehen, und da die Möglichkeiten der Grafikkarten voranschreiten, werden die Algorithmen beständig verbessert und die Approximationen nähern sich den physikalischen Strukturen der Objekte und deren Interaktion mit Licht immer weiter an.

Anders stellt sich die Situation dar, wenn die Qualität der Algorithmen einen Einfluss auf eine medizinische Diagnose bewirken könnte. In diesem Falle möchten wir diesen Einfluss früh genug erkennen, die Algorithmen, die eine Diagnosestellung ungünstig verändern könnten, ausschalten und die Algorithmen, die eine Diagnosestellung begünstigen, bevorzugt in der Medizin eingesetzt wissen. Die Qualität der Rendering-Algorithmen wird durch unterschiedliche Parameter beschrieben. Die Informatik hat gute Methoden entwickelt, die Effizienz von Algorithmen zu beurteilen und ihre Usability zu evaluieren. Wenn aber in der Medizin Bilder als Grundlage für Diagnosen verwendet werden, dann entscheiden Menschen auf Grund von interpretierbaren Merkmalen in diesen Bildern. Also müssen wir die menschliche Interpretation von Bildern empirisch auf Fragestellungen der Diagnose hin testen. Solche kontrollierten Studien mit Versuchspersonen müssen sorgsam erstellt werden. In diesem Beitrag beschäftigen wir uns mit medizinischen Bildern, die durch Volumenrendering-Algorithmen aus dreidimensionalen Datensätzen mittels Computertomografie entstanden sind.

Die Autorinnen und Autoren



Gitta Domik ist seit 1993 Professorin für Computergrafik, Visualisierung und Bildverarbeitung an der Universität Paderborn. Ihre derzeitigen Forschungsinteressen sind in den Bereichen medizinische Visualisierung und Serious Games. Sie ist ein Mitglied des ACM SIGGRAPH Education Komitees und seit 2010 am Editorial Board von „IEEE Computer Graphics and Applications“, wo sie gemeinsam mit Scott Owen das Education Department führt. Ihr Studium und ihre Promotion schloss sie an der Technischen Universität Graz ab, bevor sie 1985 für acht Jahre in den USA (Vexcel Corporation und University of Colorado at Boulder) arbeitete.



Stephan Arens ist seit 2009 wissenschaftlicher Mitarbeiter in der Fachgruppe Computergrafik, Visualisierung und Bildverarbeitung an der Universität Paderborn. Seinen Abschluss im Diplomstudiengang Informatik ergänzte er mit dem Nebenfach Medienwissenschaften. Im Rahmen seiner Promotion befasst er sich mit der Visualisierung von medizinischen Volumendaten des Herzens.



Jan Tünnermann erhielt 2010 seinen M.Sc. in Informatik mit Nebenfach Psychologie an der Universität Paderborn. Seitdem ist er dort wissenschaftlicher Mitarbeiter und Promotionsstudent in der Psychologie und Elektrotechnik. Er beschäftigt sich zum Einen mit den psychologischen Grundlagen visueller Aufmerksamkeit, und zum Anderen mit künstlichen Aufmerksamkeitssystemen für die Robotik.



Ingrid Scharlau ist seit 2007 Professorin für Kognitive Psychologie an der Universität Paderborn und zur Zeit für 3 Jahre als Professorin für Psychologie an die Universität Lüneburg beurlaubt. Ihre Forschungsschwerpunkte liegen in den Bereichen Aufmerksamkeit, Zeitwahrnehmung und Verarbeitung nicht-bewusster Information. Weitere Interessen liegen im Bereich der Verbesserung der Lehre und der interdisziplinären Zusammenarbeit. Vor ihrer Habilitation in Psychologie (Universität Bielefeld, 2005) und der Promotion in Pädagogik, Philosophie und Psychologie an der Universität Bochum (1996) schloß sie ein Studium der Psychologie in Bielefeld ab. Neben den Forschungstätigkeiten war sie am Oberstufenkolleg Bielefeld und in der Hochschuldidaktik tätig.

Ergebnisse aus Volumenrendering-Algorithmen werden z. B. als Grundlage zur Diagnose von Herz-Kreislaufkrankungen, Tumoren oder Knochenbrüchen verwendet oder dienen als dreidimensionale Vorlage bei chirurgischen Eingriffen. Volumenrendering-Algorithmen haben kein fotorealistisches Pendant in der Wirklichkeit, denn ein aufgeschnittener Mensch bei einer Herzoperation sieht völlig anders aus als der Blick ins Innere des Brustkorbes modelliert durch halbdurchsichtige Voxel. Insofern ist die Freiheit bei der Wahl der Parameter eines Volumenrendering-Algorithmus besonders groß. Bevor Kreativität, Ästhetik oder Effizienz die Parameterwahl des Volumenrendering-Algorithmus bestimmen, sollte man über die dadurch veränderte Interpretation der Ergebnisbilder Gewissheit haben.

Wir fragten uns unter anderem, ob ein teurerer Beleuchtungsalgorithmus im Volumenrendering dazu führt, dass

- die Begrenzung einer Arterie besser festgestellt werden kann, und/oder
- die Krümmung einer Arterie besser erkannt werden kann.

Wir diskutieren in diesem Beitrag beispielhaft die empirischen Studien zu diesen Fragestellungen (Kapitel 3) und nehmen allgemein auf Probleme bei empirischen Studien bezüglich der Ergebnisse von Volumenrendering-Algorithmen Bezug (Kapitel 4). Vorerst stellen wir den Stand der Forschung zur Berücksichtigung der Qualität eines Visualisierungsalgorithmus, im Besonderen des Volumenrendering, dar (Kapitel 2).

2. Stand der Forschung zur Berücksichtigung der Qualität eines Visualisierungsalgorithmus im medizinischen Anwendungsbereich

Wichtig für die Qualität eines Algorithmus sind seine Zuverlässigkeit und Effizienz. Visualisierungsalgorithmen erzeugen Bilder, die von den Menschen zuverlässig interpretiert werden sollten. Betrachten kann man diese Zuverlässigkeit aus den verschiedenen Blickwinkeln der Human-Computer-Interaction (z. B. Usability), der Visualisierung (Ausdruckskraft und Wirksamkeit, siehe nächster Absatz) oder der Anwendung (in unserem Fall der Unterstützung einer Diagnose). An dieser Stelle gehen wir aber nicht tiefer auf die Effizienz ein (wir verlangen „nur“, dass die Algorithmen in Echtzeit rendern können, um die Interaktivität sicherzustellen), sondern beschäftigen uns stattdessen mit der Zuverlässigkeit der Interpretation.

Im Bereich Human-Computer-Interaction werden kontrollierte Experimente, Beobachtungen oder Fragebögen verwendet um festzustellen, wie gut die Bedienbarkeit eines Softwareprogrammes für eine spezielle Aufgabe ist (z. B. [Pla04, FJ10]). Diese Bedienbarkeit (Usability) einer Software durch medizinisches Personal soll hier aber ebenfalls nicht im Vordergrund stehen. Visualisierungsexperten definieren Qualität in erster Linie durch Expressiveness (Ausdruckskraft) und Effectiveness (Wirksamkeit) nach einer frühen Festlegung von J. Mackinlay [Mac86]. Beispielsweise ist für Mackinlay eine Repräsentation wirksamer als eine andere Repräsentation, wenn die Interpretation einer vorher festgelegten Aufgabe dadurch schneller und genauer erfolgen kann. Beispiele dieser Art der Evaluierung, angewandt auf Volumenrendering-Algorithmen, findet man z. B. in [BALP09, LR11]. In der Medizin selbst ist die

Medical Image Perception Society¹ führend in der Untersuchung der Wechselwirkung zwischen Wahrnehmung und Bildinterpretation (z. B. [Kru10]). Dabei werden aber nicht die Ergebnisse von Visualisierungsalgorithmen untersucht, sondern 2D-Bilder.

In unseren eigenen Arbeiten zur Qualitätsbestimmung von Visualisierungsalgorithmen betrachten wir vor allem die Wirksamkeit (schnelle und genaue Interpretation). Dies ist auch im Einklang mit unseren früheren Arbeiten zur Qualitätsbestimmung von visuellen Repräsentationen von komplexen Molekülen in unterschiedlichen Darstellungsarten [VDSF96]. Zudem betrachten wir nur Ergebnisse von Volumenrendering-Algorithmen der Herzregion basierend auf Computertomografie(CT)-Daten mit dem Ziel, eine Diagnose in Bezug auf koronare Herzkrankheiten durchzuführen. Eine medizinische Diagnose einer koronaren Herzkrankheit ist eine Aufgabe, die medizinisches Wissen und Erfahrung benötigt, so dass Ärzte mit ähnlichen Erfahrungswerten unsere bevorzugten Probanden wären. Eine medizinische Diagnose auf Basis eines Bildes lässt sich, laut Krupinski [Kru10], in zwei elementarere Vorgänge unterteilen: erstens die visuelle Suche im Bild und zweitens die darauf basierende Diagnose. Der erste Vorgang ist weitgehend unabhängig vom medizinischen Wissen, während die Diagnose stark abhängig vom medizinischen Wissen und der Erfahrung ist. Wir nehmen an, dass uns diese Separierung der Vorgänge in eine perzeptuelle und eine wissensbasierte Aufgabe gelingt und untersuchen deshalb nur elementare, visuelle Teilaufgaben der Diagnose. Wir können vermuten, dass jede Verbesserung bzw. jede Verschlechterung einer visuellen Teilaufgabe auf Grund einer Veränderung des Volumenrenderings einen entsprechenden Einfluss auf die Diagnose hat. Gleichzeitig erlaubt uns diese Zweiteilung empirische Untersuchungen mit Nichtmedizinern durchzuführen.

3. Beispielhafte empirische Studien

In der Medizin verwendet man Volumenrendering-Algorithmen, um aus den Schichtbildern der CT-Aufnahmen beliebige dreidimensionale Ansichten von Patientendaten zu erhalten. Wir legen uns auf Raytracing als Volumenrendering-Algorithmus fest, bei dem Strahlen durch durchscheinende (transluzente) Voxel geschossen werden, wobei die Durchsichtigkeit der Voxel je nach Anwendung durch Transferfunktionen verändert werden kann. Man geht für diese Voxel von Schattierungseigenschaften aus, die den Beleuchtungs- und Reflexionseigenschaften von opaken oder halbdurchsichtigen Objekten der realen Welt entlehnt werden. Somit zeigen die Ergebnisse dieser Algorithmen innere Strukturen der Patientendaten, die über die Möglichkeiten der direkten Untersuchung (sogar einer invasiven) eines Patienten weit hinausgehen. Abbildung 1 zeigt zwei unterschiedliche Transferfunktionen auf dem gleichen Datensatz, wobei das linke Bild den Datensatz mit einer einfachen Schattierungsfunktion [Pho75] und das rechte Bild den Datensatz mit einer Konturenverstärkenden Schattierungsfunktion [BG07] darstellt.

Nun gilt es in empirischen Studien unterschiedliche Parameter des Volumenrenderings als Ursache für eine gute oder schlechte medizinische Diagnose zu eruieren. Die unterschiedliche Auswirkung der Parameter wird dabei durch die Genauigkeit und die Geschwindigkeit bei der Lösung von visuellen Teilaufgaben einer Diagnose gemessen.



Abbildung 1: Links Volumenrendering mit einer Transferfunktion und Phong Schattierung und rechts Volumenrendering der gleichen Daten mit einer anderen Transferfunktion und Style Shading.

Krümmung und Begrenzung von Arterien mit unterschiedlichen Beleuchtungsmodellen

In zwei Studien [DASH11] wurden zwei Schattierungsfunktionen gegenübergestellt, die schon in Abbildung 1 als unterschiedlich genannt wurden: Phong gegenüber Style Shading. Obwohl diese Veränderung der Beleuchtungssituation keine Auswirkung auf die Geometrie der Darstellung hat, scheint Style Shading die Strukturen und Texturen durch die Konturenverstärkung stärker hervorzuheben. 75 % von 28 Probanden antworteten in einer Befragung nach dem Experiment, dass sie erwarten, bei den Darstellungen mit Style-Shading besser abgeschnitten zu haben [DSAS11].

In der ersten Studie betrachteten wir die Krümmung von Koronararterien. Für die Diagnose der koronaren Herzkrankheit zeigt die Krümmung der Koronararterien einen Zusammenhang zum Schweregrad der Erkrankung: höheres Plaqueaufkommen in der Arterie steht im Zusammenhang mit stärker gekrümmten Arterien. Wir stellten die folgende Hypothese für einen entsprechenden visuellen Task in dieser Studie: „Probanden können mit Style Shading schneller und korrekter erkennen, ob eine Arterie gegenüber einer anderen stärker gekrümmt ist.“ Zur Überprüfung wurde eine Within-Subject-Studie, bei der jeder Teilnehmer an jeder Bedingung teilnimmt, durchgeführt. In unserem Fall waren die Bedingungen *Style* und *Phong* gekreuzt mit *langer Anzeigedauer* und *kurzer Anzeigedauer*, um gegebenenfalls Hinweise auf eine mögliche Interaktion der Beleuchtungsmethode und Anzeigedauer zu erhalten. Aufgabe der Probanden war es, zu bewerten, welche von jeweils zwei simultan dargebotenen Arterien stärker gekrümmt ist (Abbildung 2). Wie in

solchen Studien üblich, wurden die Durchgänge, die systematisch variierte Ausprägungen der Bedingungen enthalten, in zufälliger Reihenfolge dargeboten. Ausgewertet wurden die gemittelten Leistungen der 33 Teilnehmer. Ein *t-Test* zeigte, dass die Probanden in der Lage waren, überzufällig gut die Krümmungen zu diskriminieren. Jedoch konnte kein Unterschied bezüglich *Phong* und *Style* oder Darbietungsdauer gefunden werden (Abbildung 4 (a)).

In der zweiten Studie wurde die Wirkung der Schattierungsfunktion auf eine visuelle Suche entlang einer Arterie untersucht. Eine Hauptursache einer koronaren Herzkrankheit ist Plaque in den Koronararterien. Plaque erscheint als weißes oder dunkles Merkmal (harter gegenüber weichem Plaque) am Rande der Arterien. Volumenrendering der verrauschten Daten kann sehr ähnliche Merkmale im Hintergrund der Arterien erzeugen, deshalb ist es sehr wichtig, diese Merkmale genau zu lokalisieren: in den Koronararterien oder außerhalb der Koronararterien.

Wir stellten die folgende Hypothese für einen entsprechenden visuellen Task auf: „Probanden können mit Style-Shading schneller und korrekter erkennen, ob ein Merkmal auf einer Arterie liegt oder außerhalb.“ Das Experiment, welches entwickelt wurde, um diese Hypothese zu bestätigen, ist in Abbildung 3 dargestellt. 33 Probanden entschieden 80-mal die Frage „Liegt der Cursor auf der Arterie?“ Dazu wird zuerst eine Mittellinie der Arterie gezeigt (um die entsprechende Arterie auch für Nicht-Mediziner eindeutig zu definieren) und anschließend entweder ein Phong- oder Style-Bild präsentiert. Wir konnten einen signifikanten Unterschied zugunsten von Style-Shading finden (Abbildung 4 (b)).

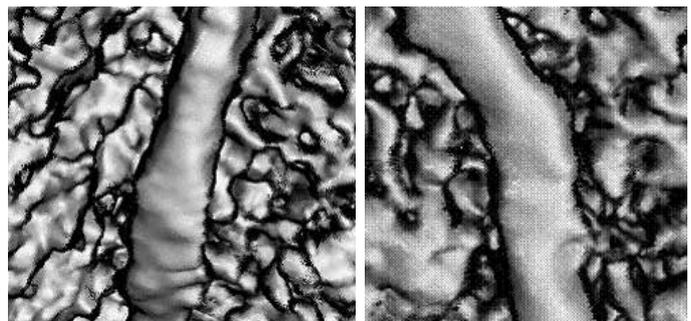
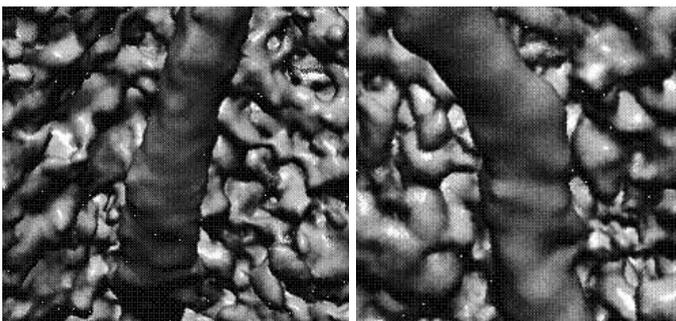


Abbildung 2: Visueller Task Krümmung „Welche Arterie (links' oder ,rechts') ist stärker gekrümmt?“ Links das Paar mit Phong-Shading, rechts dasselbe Paar mit Style-Shading.

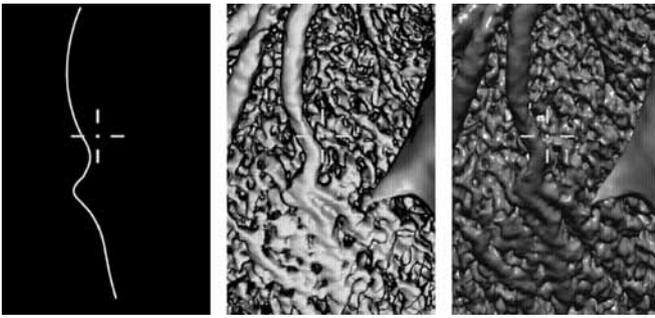


Abbildung 3: Visuelle Task „Liegt der Cursor auf der Arterie?“
Zwei Bilder werden hintereinander gezeigt: Zuerst das linke Bild (0.8 Sekunden), um auf den Verlauf der Arterie hinzuweisen. Danach wird entweder ein Style (Mitte) oder Phong (rechts)-schattiertes Bild für 100 oder 150 ms angezeigt. Der Proband entscheidet, ob der Cursor auf oder neben die Arterie zeigt.

Die Ergebnisse zeigen Schwierigkeiten der empirischen Arbeit: Es kann sich ein tendenzieller Mittelwertunterschied zugunsten Style auch in der Krümmungsaufgabe andeuten, jedoch ist das Ergebnis nicht signifikant. Andererseits heißt die Tatsache, dass wir dort keinen Unterschied finden, keinesfalls, dass es ihn nicht gibt: Ob ein Ergebnis signifikant wird, hängt von einer Reihe von Faktoren ab, beispielsweise der Größe des tatsächlichen Effektes und der Stichprobengröße.

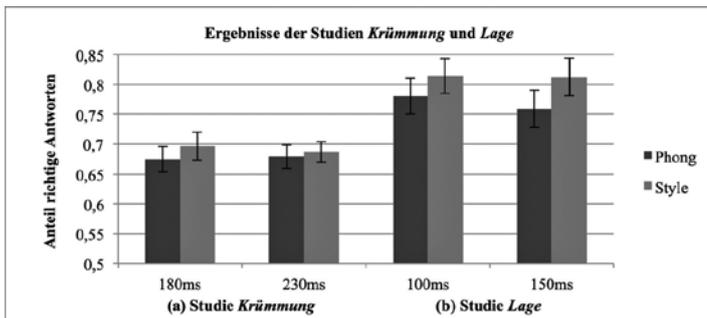


Abbildung 4: Tendenzielle Mittelwertunterschiede zugunsten von Style. Die Ergebnisse für die Krümmungsaufgabe sind jedoch nicht signifikant.

In den beschriebenen Studien interessierten wir uns für die Beleuchtung „echter“ CT-Daten. Dies bringt weitere Schwierigkeiten mit sich: Es gibt keine gänzlich verlässlichen Referenzen für die wahre Krümmung. So war es notwendig, eine vereindeutigende Vorauswahl für die Bildpaare zu treffen, bei der wir nicht-eindeutige Krümmungen ausschließen mussten. Gerade diese jedoch wären notwendig, um auch kleine Effekte finden zu können. Im Folgenden betrachten wir weitere Herausforderungen der empirischen Evaluierung von Visualisierungen.

4. Herausforderungen und Schwierigkeiten in Studien

Studien dieser Art sind häufig dadurch gekennzeichnet, dass der subjektive Eindruck bezüglich der Schwierigkeit der Aufgabe oder der Eignung des Bildmaterials stark von der tatsächlichen Leistung abweichen kann. Das ist aus psychologischer Sicht inte-

ressant und unterstreicht die Notwendigkeit solcher Forschung. Strukturelle Probleme dieser Art wurden vornehmlich bei Gedächtnisaufgaben untersucht, aber sind auch bezüglich der Wahrnehmung bekannt. So können Menschen überzufällig gut ein entferntes Symbol erraten, obwohl sie es nach subjektivem Eindruck nicht erkennen können [Sid98]. Die Einschätzung der Güte einer zur Diagnose herangezogenen Darstellung muss also nicht unbedingt mit der tatsächlichen Diagnoseleistung übereinstimmen. Insbesondere die Beschreibung von Abläufen der eigenen Informationsverarbeitung gelingt Menschen sehr schlecht. Wie oben erwähnt, gaben Probanden Style-Shading den Vorzug für die Krümmungsaufgabe; jedoch konnten wir nicht bestätigen, dass dies tatsächlich besser geeignet ist. Dies unterstreicht, dass Angaben von Benutzerinnen und Benutzern nicht ausreichen und – insbesondere bei solchen kritischen Aufgaben – handfeste empirische Belege erforderlich sind.

Zumeist müssen solche Belege durch Experimente erbracht werden. Für deren Konzeption, Durchführung und Auswertung wird in der Psychologie streng wissenschaftlich vorgegangen: Einzelne Faktoren werden isoliert und variiert (unabhängige Variable; UV) und so ihr Einfluss auf die Leistung oder andere abhängige Variablen (AV) erfasst. Dies ist keineswegs trivial, da mögliche Unterschiede oft klein sind, die Varianz zwischen unterschiedlichen Versuchspersonen jedoch häufig groß ist. So begründet sich auch die Aufteilung in unterschiedliche Darstellungsdauern in den oben beschriebenen Versuchen damit, dass ein Vorversuch einen erheblichen Einfluss dieser nahelegte, der Hauptversuch, mit ausreichender Anzahl an Versuchspersonen für eine statistisch gesicherte Aussage, diesen jedoch nicht bestätigen konnte. Anders als in anderen Naturwissenschaften spielt in der Psychologie unerwünschte Datenvarianz eine sehr erhebliche Rolle. Sie erfordert, dass Experimente sorgfältig geplant werden, da sie so hoch sein kann, dass tatsächlich existierende Effekte der UV statistisch nicht mehr nachgewiesen werden können.

Aufgabe des Experimentators ist es, die auf die experimentelle Manipulation zurückgehende sogenannte Primärvarianz zu maximieren, etwa, indem extreme Stufen gewählt werden (so etwa die stärkste Approximationsstufe), Störfaktoren kontrolliert und die Zufallsvarianz minimiert wird, zum Beispiel durch stärkere Standardisierung der Untersuchungssituation oder hinreichend zuverlässige Messinstrumente [Ker73].

Störfaktoren sind Variablen, die unabsichtlich mit variiert werden und, so eine alternative Erklärung für einen ermittelten Effekt sein können. Ein Beispiel hierfür findet sich in einem Versuch, den wir durchgeführt haben, in dem Probanden angeben mussten, welcher von zwei Punkten, die auf der Visualisierung eines CT-Herzdatensatzes markiert wurden, weiter vorn liegt (siehe Abbildung 5). Auch hier sollte untersucht werden, ob diese Aufgabe besser mit einem globalen Beleuchtungsalgorithmus oder lokaler Phong-Schattierung erledigt werden kann. Im globalen Beleuchtungsmodell schien jedoch eine bestimmte Art von Schatten (sehr weiche) dazu zu führen, dass Punkte an diesen Orten als weiter hinten wahrgenommen werden, als sie eigentlich sind. So erzeugen unterschiedliche Beleuchtungsmethoden oft auch unterschiedliche Helligkeitskontraste. Findet man einen Leistungsunterschied zwischen den Beleuchtungsmethoden, ist zu prüfen, ob er tatsächlich auf den Algorithmus oder lediglich auf den Kontrast zurückgeht.

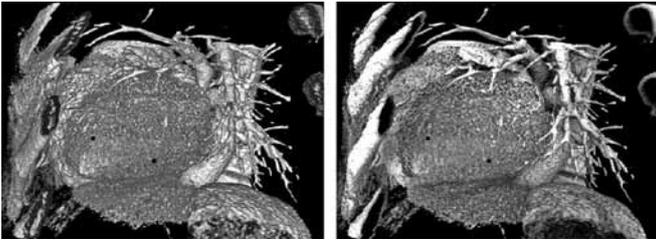


Abbildung 5: Visuelle Studie Tiefenwahrnehmung „Welcher der beiden roten Punkte liegt näher am Betrachter?“ Links eine Visualisierung eines Herzensatzes mit Phong-Shading, rechts eine Visualisierung mit globaler Beleuchtung.

Üblicherweise werden Störeinflüsse eliminiert, indem man sie in den Bedingungen konstant hält oder so randomisiert, dass sie keinen systematischen Einfluss auf die UV oder AV haben. In unseren Beispielen ist das jedoch schwierig: Ist der Schatten ein Störfaktor oder ein integraler Bestandteil der Beleuchtungsmethode? Wird er, wenn er mit einer härteren Abgrenzung gerendert wird, besser als Schatten erkennbar und der störende Einfluss verschwindet? Wenn der pure Helligkeitskontrast für Phong und Style konstant gehalten werden soll, muss ein globales oder ein lokales Kontrastmaß verwendet werden? Inwieweit hängt dies von der Aufgabe ab?

Neben diesen spezifischen Schwierigkeiten findet man in verwandten Arbeiten häufig auch grundlegende Probleme: Die Anzahl an Versuchspersonen ist oft zu gering oder es werden zu wenige Wiederholungen einer Bedingung realisiert, um aussagekräftige Statistik zu betreiben. Grundsätzlich ist an manchen Studien zu bemängeln, dass Expertenwissen zur sorgfältigen Durchführung eines Experiments fehlt, sodass sich schon in der Planung des Experimentes Fehler einschleichen, die auch bei korrekter mathematischer Auswertung zu Fehlern in den Ergebnissen führen. Davon unabhängig werden Ergebnisse gelegentlich fehlinterpretiert. Das ist umso bedauerlicher, da viele dieser empirischen Studien im universitären Umfeld durchgeführt werden, wo neben der medizinischen Expertise auch die entsprechende Expertise in der Psychologie verfügbar ist.

5. Zusammenfassung und Ausblick

Der vorangegangene Abschnitt verdeutlicht bereits die Richtung für zukünftige Arbeiten: Die Auswirkungen von Beleuchtungsalgorithmen müssen noch genauer betrachtet werden. Im zuletzt beschriebenen Beispiel müssten die Schatteneigenschaften gezielt variiert werden, um den Einfluss zu verstehen. Für viele andere Studien (nicht nur unsere) muss überprüft werden, inwieweit Helligkeitskontraste Leistungsunterschiede erklären. Diese stehen stellvertretend für eine ganze Reihe an Parametrisierungen der Methoden und es muss untersucht werden, inwiefern die Methode an sich oder bestimmte Einstellungen positiv oder negativ zu bewerten sind. Schließlich bleibt ein weiteres Thema, das nicht nur im vorliegenden Text wenig behandelt wurde: Die Ergebnisse von Wahrnehmungsexperimenten in diesem Kontext liefern oft nur Erkenntnisse, die eine sehr spezifische Aufgabe betreffen, und oft kann nur wenig generalisiert werden. Daher ist es wichtig zu verifizieren, dass positive Effekte sich auch auf die wirkliche Aufgabe in der Anwendung übertragen. Zum Beispiel

ist möglich, dass sich eine Darstellung positiv auf das Finden eines Ziels auswirkt, jedoch negativ auf dessen Klassifizierung. Je nachdem, welche der beiden Aufgaben in der Praxis mehr Unterstützung bedarf (oder welche kritischer ist), kann sich die Darstellung positiv oder negativ auf die Gesamtaufgabe auswirken.

Zusammenfassend kann gesagt werden, dass das empirische Evaluieren zur Wahrnehmung von Visualisierungen mit Methoden, die psychologischen Standards genügen, wichtig und der einzige Weg zu gesicherten Erkenntnissen ist. Wir haben deutlich gemacht, dass die Notwendigkeit oft erst durch den Prozess selbst erkennbar wird, da Introspektion der Anwenderinnen und Anwender und intuitiv wahrgenommene Qualität von Visualisierungen oft nicht mit der messbaren Leistung übereinstimmen. Diese Art der Evaluierung ist oft schwierig und aufwendig im Verhältnis zur gewonnenen Erkenntnis. In Alltagsanwendungen, in denen sowieso häufig subjektiver Eindruck und Akzeptanz durch die Nutzer die ausschlaggebenden Kriterien sind, sind solche Methoden nicht notwendigerweise lohnenswert. In kritischen Bereichen, wie im hier diskutierten medizinischen Kontext, sind sie jedoch unumgänglich, und weitere Forschung ist dringend erforderlich.

Anmerkungen

1 <http://www.mips.ws/>

Danksagung

Für die medizinische Expertise und Vorlage von Daten danken wir unseren langjährigen Kooperationspartnern, Dipl.-Ing. R. Weise, Dr. H. Fricke und Prof. Dr. med. W. Burcher des Herz- und Diabeteszentrums Bad Oeynhausen. Weiterhin gelten die Danksagungen aus den hier zitierten Studien nach [DASH11] und an die Studierenden Ph. Daraio, F. Isenberg und P. Markwart.

Referenzen

- [BALP09] Baer A., Adler F., Lenz D., Preim B.: Perception-based evaluation of emphasis techniques used in 3d medical visualization. In VMV (2009), pp. 295–304.
- [BG07] Bruckner S., Gröller M. E.: Style transfer functions for illustrative volume rendering. Computer Graphics Forum 26, 3 (Sept. 2007), 715–724. Eurographics 2007 3rd Best Paper Award.
- [DASH11] Domik G., Arens S., Scharlau I., Hilkenmeier F.: How Useful is Computer Graphics for Medical Diagnoses? In Emerging Technologies for Medical Diagnosis and Therapy, 41. Jahrestagung der Deutschen Gesellschaft für Informatik GI-2011.
- [DSAS11] Domik G., Steffen F., Arens S., Scharlau I.: Usefulness of style transfer functions in medical diagnosis. In SIGGRAPH Posters (2011), p. 89.
- [FJ10] Forsell C., Johansson J.: An heuristic set for evaluation in information visualization. In Proceedings of the International Conference on Advanced Visual Interfaces (New York, NY, USA, 2010), AVI '10, ACM, pp. 199–206.
- [Ker73] Kerlinger F.: Foundations of behavioral research. New York: Holt, Reinhart and Winston. 1973.
- [Kru10] Krupinski E.: Current perspectives in medical image perception. Attention, Perception, & Psychophysics 72 (2010), 1205–1217. 10.3758/APP.72.5.1205.

- [LR11] Lindemann F., Ropinski T.: About the Influence of Illumination Models on Image Comprehension in Direct Volume Rendering. IEEE TVCG (Vis Proceedings) 17, 12 (2011), 1922–1931.
- [Mac86] Mackinlay J.: Automating the design of graphical presentations of relational information. ACM Trans. Graph. 5, 2 (Apr. 1986), 110–141.
- [Pho75] Phong B. T.: Illumination for computer generated pictures. CACM 18, 6 (June 1975), 311–317.
- [Pla04] Plaisant C.: The challenge of information visualization evaluation. In Proceedings of the working conference on Advanced Visual Interfaces (New York, NY, USA, 2004), AVI '04, ACM, pp. 109–116.
- [Sid98] Sidis B.: The Psychology of Suggestion: A Research into the Subconscious Nature of Man and Society. New York: D. Appleton & Co., 1898.
- [VDSF96] Volbracht S., Domik G., Shahrabaki K., Fels G.: An Experimental Comparison of 3D Display Modes. Proceedings of IEEE Visualization '96, IEEE Computer Society Press. Late Breaking Hot Topics Papers, 8-11.

*erschienen in der Fiff-Kommunikation,
herausgegeben von Fiff e.V. - ISSN 0938-3476
www.fiff.de*