Interaktives Streaming von hochaufgelösten 360°-Panoramen

Carsten Grünheit, Aljoscha Smolic und Thomas Wiegand
Heinrich-Hertz-Institut Berlin (HHI)
Bildsignalverarbeitung
Einsteinufer 37, 10587 Berlin, Germany
{gruenheit/smolic/wiegand}@hhi.de

Kurzfassung
Ein Server/Client-System für das interaktive Streaming von hochaufgelösten 360°-Panoramen wird vorgestellt. Die Panoramen sind in MPEG-4-BIFS-Szenen eingebettet, welche über das Internet übertragen werden. Die Panoramen werden dazu in Teilbilder zerlegt, vorcodiert und auf einem Server abgelegt. Auf der Client-Seite kann der Benutzer über einen MPEG-4-Player durch das Panorama navigieren. Um ein flüssiges Navigieren auch bei großen Datenmengen zu ermöglichen, werden die Interaktionen des Benutzers mit der Szene ausgewertet und nur jeweils die Daten übertragen, die notwendig sind, um die aktuelle Bildschirmansicht zu erzeugen. Aufgrund der übertragungsbedingten Verzögerungszeiten ist hierzu ein Pre-Fetching-Mechanismus notwendig, der sicherstellt, dass Bilddaten, die mit einer gewissen Wahrscheinlichkeit für die Bildschirmansicht benötigt werden, rechtzeitig angefordert werden.

1. Einleitung

In den letzten Jahren hat es im Bereich der Computertechnik große Fortschritte gegeben. Heutige PCs bieten gute Voraussetzungen im Bereich der Rechenleistung von CPUs und Grafik-Prozessoren und der Verfügbarkeit von Arbeitsspeicher und Festplattenplatz, um auch kom-

plexe Multimedia-Anwendungen realisieren zu können. Diese Entwicklung wird begleitet von einer immer größer werdenden Verbreitung des Internets. Hierbei nimmt auch die Zahl der Nutzer mit einer Hochgeschwindigkeitsverbindung ständig weiter zu.

Als Konsequenz dieses technologischen Fortschritts wird es zunehmend interessant, neue, komplexe Multimedia-Dienste anzubieten. Eine Vision dabei ist es, einem Benutzer die Möglichkeit zu geben, sich frei in dreidimensionalen virtuellen Szenen bewegen zu können, die dem User einen immersiven Eindruck vermitteln, d.h. ihm das Gefühl vermitteln, tatsächlich mit einer realen Szene zu interagieren. Ließen sich die dafür notwendigen Daten im Internet vielen Benutzern zugänglich machen, eröffneten sich Möglichkeiten für neue Dienste, z.B. neue Formen des virtuellen Tourismus. Der Benutzer könnte sich am eigenen PC virtuell in Landschaften, Städte, zu Sehenswürdigkeiten usw. begeben.

Heutige Möglichkeiten des virtuellen Tourismus beschränken sich meist auf das Konsumieren von schriftlichen, Audio- und Einzelbild-Informationen. Fotografien bieten dabei natürlich nur einen festgelegten Blickpunkt, haben keine zeitliche Dimension und erlauben keine Interaktion mit der Szene. Auf Kosten einer erhöhten Datenrate können auch Videosequenzen angeboten werden.

Ganz andere Anwendungen, die sogar ein gemeinsames Interagieren mehrerer Benutzer mit einer dreidimensionalen Szene erlauben, stellen z. B. Netzwerk-basierte Computerspiele dar. Dabei können sich Benutzer frei in mit Szenenbeschreibungssprachen wie der Virtual Reality Modeling Language (VRML) komplett computer-generierten virtuellen Umgebungen bewegen. Da mit vertretbarem Aufwand erzeugte Szenen jedoch immer weniger Details aufweisen als die Realität, entsteht auch hier kein immersiver Eindruck.

In den letzten Jahren wurden neue Technologien zur Akquisition und Visualisierung von dreidimensionalen Umgebungen entwickelt, die auf echten photorealistischen Aufnahmen beruhen. Als Beispiele für diese neuen Image-Based Rendering (IBR) –Verfahren, bei denen Einzelbilder als Abtastwerte der plenoptischen Funktion [3] aufgefasst werden, sind insbesondere Light-Field Rendering [1] und die Concentric Mosaics Methoden [2] zu nennen. Hierbei wird reales Bildmaterial verwendet. Unabhängig davon, welches Verfahren im Einzelnen benutzt wird, ist der Grad des immersiven Eindrucks abhängig von der Menge der verwendeten Bilddaten. Das bedeutet, je überzeugender der visuelle Eindruck einer Szene wird, desto schwieriger wird es, dies über ein Netzwerk Benutzern zugänglich zu machen. Werden statt Standbildern Videosequenzen zum Generieren einer Szene verwendet (Video-Based Rendering (VBR)), wird das Übertragungsproblem weiter vergrößert, z.B bei Immersive Panoramic Video [10].

Im Internet gibt es bereits frei verfügbare Anwendungen, die es erlauben, auf realem Bildmaterial basierende Szenen anzuzeigen. Das verbreitete QuickTimeVR® System [4] erlaubt Zoom und Rotation in zylindrischen oder kubischen Rundumansichten. Dafür müssen alle Bilddaten zuvor komplett heruntergeladen werden. Um dabei aber eine vertretbare Reaktionszeit des Systems realisieren zu können, ist die visuelle Qualität begrenzt (kleine Bilder, geringe Auflösung, Aliasing bei Bewegung).

Das Ziel dieser Forschungsarbeit ist es, die Übertragung und Interaktion mit photorealistischen 3D-Welten über das Internet zu ermöglichen, die auf sehr großen Bilddatenmengen beruhen, die einen initialen Download in akzeptabler Zeit nicht ermöglichen. Das interaktive Streaming von hochaufgelösten 360°-Panoramabildern stellt einen Schritt auf dem Weg dahin dar.

2. Interaktives Streaming

Der von uns gewählte Ansatz, um bildbasierte Verfahren in Internet-Diensten verwenden zu können, beruht darauf, die Daten interaktiv von einem Server zu streamen. Es sollen nur genau die Daten zum Client übertragen werden, die erforderlich sind, um die aktuelle Ansicht der Szene rendern zu können und es dem User erlauben, darin zu navigieren. Um dies zu ermöglichen, müssen die Navigationsentscheidungen des Benutzers für die Auswahl der aktuell zu übertragenden Bilddaten ausgewertet werden.

Beim Aufbau eines solchen Systems stellen sich grundsätzliche Fragen bzgl. einer geeigneten Szenenrepräsentanz, des zu wählenden Codierverfahrens, des Streamingverfahrens und eines geeigneten Players, um die Szene anzuzeigen und darin navigieren zu können.

Da bei diesem Szenario von sehr großen Datenmengen ausgegangen wird, ist ein effizientes Kompressionsverfahren notwendig. Um ein für viele Benutzer zugängliches System aufbauen zu können, liegen die Bilddaten vorcodiert auf einem Server. Daher können die Navigationsentscheidungen des Benutzers nicht bei der Codierung berücksichtigt werden. Es steht weder eine bekannte zeitliche Anordnung wie bei Videosequenzen zur Verfügung, noch ist der Pfad der Bewegung bekannt. Ganz im Gegenteil soll ein wahlfreier Zugriff auf die vorcodierten Daten ermöglicht werden, um dem Benutzer ein Maximum an Bewegungsfreiheit zu bieten. Daher sind auch prädiktive Codierverfahren nur eingeschränkt anwendbar.

Auch die notwendige Auflösung ist keine feste Größe mehr, sondern wird vielmehr abhängig von dem gewählten Beobachtungspunkt und Zoom. Ein fertiges System sollte daher eine Skalierbarkeit von Codierung und Übertragung unterstützen.

Um eine störungsfreie Navigation zu ermöglichen, ist es insbesondere notwendig, die auftretenden Verzögerungszeiten zu berücksichtigen. Allein die Roundtrip-Time einer Nachricht kann im Internet bereits mit 100 ms geschätzt werden. Zusammen mit den Hardware-abhängigen Verarbeitungsverzögerungen bildet die Gesamtverzögerung ein Problem, das es nicht erlaubt, Daten direkt nach dem Bedarf des Clients zu übertragen. Stattdessen ist es bei interaktivem Streaming erforderlich, die Daten nach einer Pre-Fetching-Strategie vom Server anzufordern, die wahrscheinlich demnächst benötigt werden, um die dann aktuelle Ansicht der Szene darstellen zu können.

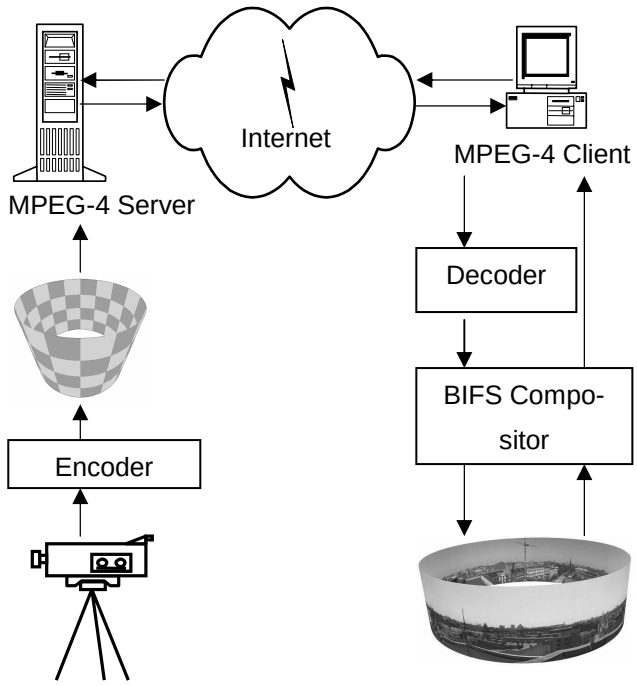

Abbildung 1: Prinzipbild eines interaktiven Streamingsystems

3. Streaming von Panorama-Szenen

Abbildung 1 zeigt das Prinzip des von uns entworfenen kompletten interaktiven Streamingsystems. Es werden Szenen mit hochaufgelösten 360°-Panoramabildern erzeugt, die ein Benutzer auf der Client-Seite in-

teraktiv betrachten kann. Es wurden hochaufgelöste Panoramen als Beispiel ausgewählt, da auf diese Weise relativ leicht Szenen mit sehr großen Bilddatenmengen erzeugt werden können. Die Panoramabilder werden mit Spezialkameras oder mit gewöhnlichen Videokameras in Kombination mit robuster globaler Bewegungsschätzung und Mosaicing Tools erzeugt [5], [6], [7]. Dabei werden alle Bilder einer Videosequenz transformiert und zu einem einzigen großen Bild zusammengesetzt.

Die von uns entwickelte Szenenrepräsentation besteht neben den Bilddaten aus Szenenbeschreibungsinformationen in MPEG-4 BIFS (Binary Format for Scenes), das in wesentlichen Teilen auf VRML basiert. BIFS wurde zur Darstellung multimedialer Daten entwickelt, ist sowohl für Audio, visuelle und Objektdaten verwendbar und eignet sich daher sehr gut für die vorgestellte Anwendung. Des Weiteren kann der Szenenaufbau leicht verändert oder erweitert werden, ohne Änderungen am Gesamtsystem vornehmen zu müssen.

Dargestellt wird die Szene auf der Client-Seite mit dem HHI MPEG-4 Player [8]. Mit diesem ist es möglich, frei in der Szene zu navigieren. Es ist auch möglich, sich in dem Zylinder zu drehen, zu zoomen und auch darin zu laufen, auch wenn in dem von uns gewählten Zylinder-Szenario der einzige, völlig verzerrungsfreie Beobachtungsstandort auf der Zylinderachse mit radialer Blickrichtung auf die Innenseite des Zylinders ist.

Zur Zeit erfolgt das Streaming der Daten mit einer am HHI entwickelten MPEG-4 Client/Server-Architektur [8], die den MPEG Delivery Multimedia Integration Framework (DMIF) implementiert. Damit ist es möglich, Anwendungen unabhängig von den Details der verwendeten Übertragungstechnologie zu entwerfen. Die Verbindung hierzu wird über das DMIF Application Interface (DAI) hergestellt. In der vorgestellten Anwendung wird das System zur Echtzeit-Übertragung von Multimedia-Inhalten über das Internet verwendet.

Weder der Player, noch das Übertragungssystem schreiben die Verwendung bestimmter Codecs für die Einzelbildcodierung vor. Im Moment wird JPEG für die Codierung der Panorama-Bilddaten verwendet. Dieses Verfahren bietet eine gute Kompression und sollte von jedem System unterstützt werden, das ISO/IEC 14496-1 (MPEG-4 Systems) implementiert. Dieser Codec wird demnächst durch JPEG2000 ersetzt werden, der erweiterte, für die angestrebte Anwendung vorteilhafte Eigenschaften bietet [9], wie den wahlfreien Zugriff auf die codierten Daten sowohl auf der Tile-Ebene als auch auf der Precinct-Ebene, bei der auf Gruppen von Codeblöcken zugegriffen werden kann. Skalierbarkeit ist bereits vorgesehen, und der Bitstrom kann nach steigender Auflösung, Qualitätsebenen, Anzahl der Komponenten und räumlicher Position organisiert werden.

Szenenrepräsentation

Jede Grafikkarte hat nur eine begrenzte Größe von Texturspeicher zur Verfügung. Daher ist es notwendig, die Komplexität der zu rendernden Szene und die Menge der darzustellenden Bilddaten zu steuern. Die von uns entwickelte Technik ermöglicht dies unter ausschließlicher Verwendung von MPEG-4 BIFS-Knoten. Hierzu wird das Panoramabild in kleine Teilbilder (< 512x512 Pixel) zerlegt, die separat codiert werden und dem jeweiligen entsprechenden Abschnitt des Zylinders zugeordnet werden, der unsichtbar für den Betrachter in einzelne Objekte zerlegt wird.

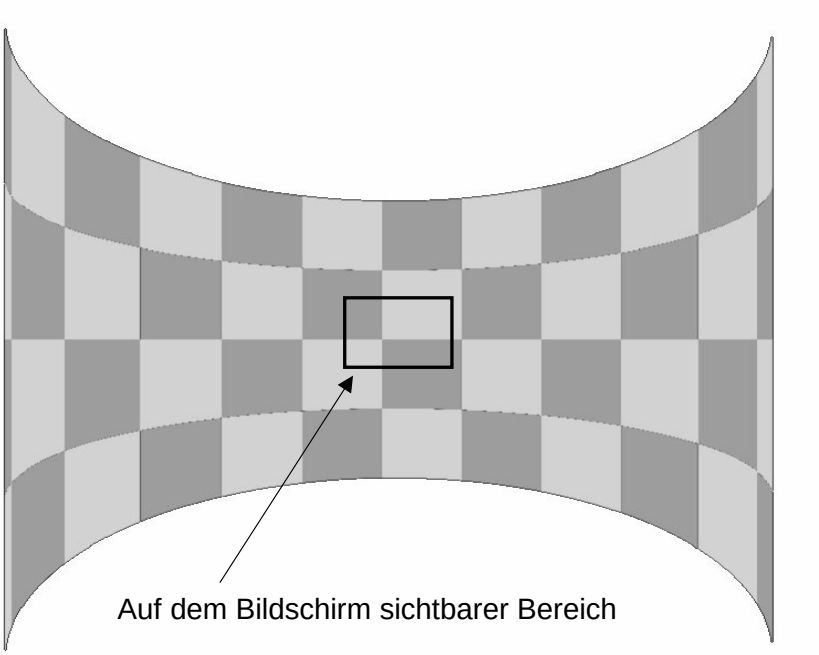

Abbildung 2: Zylindrisches Panorama, zerlegt in Segmente

Mit dieser prinzipiell einfachen Methode wird es dem BIFS-Compositor, der das Kernstück des verwendeten Players bildet, ermöglicht, einzelne Segmente des Zylinders in den zu rendernden Teil der Szene zu laden und sie selbständig auch wieder zu entfernen, sobald diese den sichtbaren Bereich wieder verlassen haben.

Der Mechanismus, mit dem das Laden und das Entfernen realisiert wird, dient auch dazu, die benötigten Bilddaten vom Server anzufordern, bevor der Benutzer die betroffene Region des Zylinders sieht. Im Kern basiert das implementierte Verfahren auf der Verwendung des BIFS Visibility-Sensor-Knotens, mit dem die Sichtbarkeitsänderungen einer den Objekten zugeordneten Bounding Box verfolgt werden können, wenn der User durch die Szene navigiert. Jedem Zylindersegment ist ein dieses umschließender Sichtbarkeitssensor zugeordnet, vgl. Abbildung 3. Tritt der Sensor in den sichtbaren Bereich der Szene ein, wird ein Ereignis generiert, das in eine Anforderung umgesetzt und zum Server geschickt wird, der dann die Elementarströme mit den zugeordneten Daten zum

Client schickt. Sobald der Sensor sich nicht mehr mit dem sichtbaren Bereich überschneidet, entfernt der BIFS Compositor die Texturdaten aus der zu rendernden Szene, so dass die Summe aller Zylindersegmente mit Teilbildern, die Einfluss auf die Rendering Performance haben, begrenzt ist durch die Anzahl der aktivierten Visibility-Sensoren.

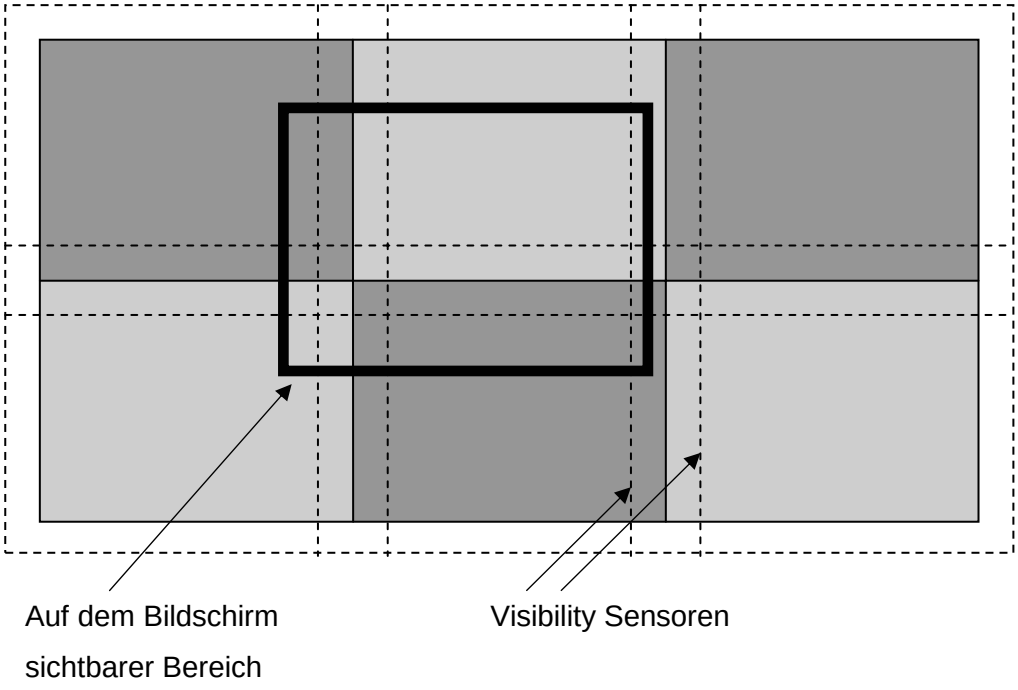

Abbildung 3: Zylindersegmente mit umschließenden Sichtbarkeitssensoren

Ein Pre-Fetching von Bilddaten, d.h. deren vorausschauendes Anfordern für die Bildschirmansicht, lässt sich am einfachsten durch eine Überdimensionierung der Bounding Boxen der Sensoren erreichen. In Abbildung 3 werden z.B. alle sechs Teilbilder vom Server angefordert, obwohl die beiden rechten noch nicht im sichtbaren Bereich liegen.

Es ist erforderlich, einen Kompromiss zwischen der Dimension der Sensoren und der angestrebten Rendering Performance zu finden: Je größer die Sensoren gewählt werden, desto mehr Bilddaten werden in die zu rendernde Szene geladen. Dies kann zu einer schlechteren Performance führen, die sich z.B. durch ruckhafte Bewegungen äußert. Aller-

dings garantiert eine ausreichende Größe der Sensoren, dass die Bilddaten tatsächlich verfügbar sind, wenn der User die zugehörige Region auf dem Zylinder aufdeckt, auch wenn die Navigationsgeschwindigkeit hoch sein sollte.

Die Zeit, die zur Verfügung steht, um die Daten anzufordern, zu übertragen und zu verarbeiten, wird daher bestimmt durch das Größenverhältnis der Zylindersegmente zu den sie umschließenden Sensoren. Daher wird ohne Verwendung einer skalierbaren Codierung die mögliche Geschwindigkeit der Navigation begrenzt, wenn keine störenden Artefakte sichtbar werden sollen. Überschreitet man diese maximale Geschwindigkeit, ist es möglich, Regionen des Zylinders aufzudecken, für die noch keine Texturen vorliegen. Sobald sie dann verfügbar sind, werden diese verspätet in ihre bis dahin leeren Positionen auf dem Zylinder geschrieben.

Die Zeitspanne zwischen der Anforderung und dem Rendern der Teilbilddaten ist abhängig von der zur Verfügung stehenden Bandbreite und der Verarbeitungsszeit für die Elementarströme, in denen die codierten Daten transportiert werden. Die Rendering Performance hängt aber auch insbesondere von der Texturpuffergröße der Grafikkarte ab, in der die anzuzeigenden Werte abgelegt werden, da bei einem Überlauf die Daten im Hauptspeicher des PC abgelegt werden.

Unabhängig von den gegebenen Netzwerk- und Hardware-Voraussetzungen kann die Performance des Systems dadurch beeinflusst werden, dass Einschränkungen in die Navigationsfreiheit des Benutzers gemacht werden. Die Obergrenze des Zoom-Bereiches, d.h. der maximale Öffnungswinkel der virtuellen Kamera, legt die Anzahl der maximal notwendigen Zylindersegmente und damit die maximale Bitrate und Menge der zu rendernden Bilddaten fest. Außerdem besteht die Möglichkeit, die notwendige Übertragungsbandbreite zu begrenzen, in-

dem die maximale Navigationsgeschwindigkeit heruntergesetzt wird. Durch ein Einstellen dieser Parameter kann das System für unterschiedliche Hardware- und Netzwerkumgebungen optimiert werden.
Da zur Zeit die Szenenrepräsentation vollständig mit Standard-BIFS-Knoten realisiert ist, genügt das gesamte interaktive Streamingsystem für hochaufgelöste Panoramen dem MPEG-4 Standard.

5. Beispiele

Unser System gestattet das interaktive Streaming von Panoramaansichten von nahezu unbegrenzter Größe. Es wurde getestet mit Bildern der Größe bis 13200x2600 Pixel, die unkomprimiert einen Speicherplatz von etwa 100 MB erfordern. Wir arbeiten daran, das System auch mit Bildern der Größe 10000x60000 Pixel zu testen, was etwa 1,7 GB unkomprimierter 24-Bit RGB-Daten entspricht, bei weitem zuviel, um die Daten im Vorhinein komplett herunterzuladen.

Abbildung 4: Komplettansicht eines zylindrischen Panoramabilds

Abbildung 4 zeigt ein kleineres Panoramabild (3700x400 Pixel) in der Komplettansicht, das mit Mosaicing erzeugt wurde. Abbildung 5 zeigt beispielhaft zwei Screenshots von Ansichten auf dieses Panorama. Um von der linken Ansicht zur rechten zu gelangen, wurde rotiert und dann herangezoomt.

Abbildung 5: Einzelne Ansichten eines Panoramazylinders (3700 x 400 Pixel), Berlin

Abbildung 6: Einzelne Ansichten eines Panoramazylinders (13200 x 2600 Pixel), Auckland[1]

Abbildung 6 zeigt weitere Screenshots einer 13200x2600 Pixel großen 360°-Panorama-Szene, aufgenommen auf der Hafenbrücke in Auckland,

[1] Bildmaterial zur Verfügung gestellt durch DLR-Institut für Weltraumsensorik

Neuseeland. In unserer Szenenrepräsentation ist für dieses Panorama ein maximaler Öffnungswinkel der virtuellen Kamera von 63° möglich, ohne dass die Zylindergrenzen sichtbar werden. Bei einer Unterteilung des Panoramas in 156 Teilbilder, einem sinnvoll gewählten Öffnungswinkel von 28°, bei dem bis zu 4x4 Teilbilder gleichzeitig sichtbar sind, und einem Größenverhältnis von 1,4:1 von Sensor und umschlossenen Zylindersegment, ist es möglich, in der Panoramaszene mit einer Geschwindigkeit von ca. 5°/ sec bei akzeptabler Rendering Performance und ohne Fehler aufgrund sichtbar verspätet angezeigter Teilbilder zu rotieren. Diese Ergebnisse werden erzielt bei fehlerfreier Übertragung und einer verfügbaren Bandbreite von max. 1 MBit/sec pro Elementarstrom. Verkleinert man den Öffnungswinkel oder reduziert die Rotationsgeschwindigkeit etwas, ist eine visuell völlig einwandfreie Navigation möglich. Vergrößert man die Sensoren, kann schneller rotiert werden ohne „verspätete“ Texturen, allerdings auf Kosten einer weniger weichen Bewegung. Das System wurde auf einem PC mit 2x2 GHz Intel® Xeon Prozessor und Nvidia® Quadro2 Pro Grafik-Chipsatz getestet. Bei der Verwendung einer 1,3 GHz Intel® Pentium3 Mobile CPU und einem Nvidia® GeForce2 Go Grafik-Chipsatz werden vergleichbare Ergebnisse bei einer Rotationsgeschwindigkeit von etwa 3,3°/sec erreicht.

6. Zusammenfassung

Es wurde ein neues System vorgestellt, mit dem hochaufgelöste Panoramaansichten über das Internet übertragen werden können. Dabei bietet das interaktive Streaming-System die Möglichkeit, durch die Szene zu navigieren, ohne zuvor alle Bilddaten vom Server herunterladen zu müssen. Stattdessen werden immer gerade die Daten auf Anforderung übertragen, die zum Erzeugen der aktuellen Bildschirmansicht und für

eine störungsfreie Navigation durch die Szene benötigt werden. Das System ist kompatibel mit dem MPEG-4 Standard.

Als nächstes wird die Integration von JPEG2000 realisiert werden, um Skalierbarkeit und einen verbesserten wahlfreien Zugriff auf die Bilddaten zu ermöglichen. Des Weiteren werden komplexere Pre-Fetching-Strategien und das Verhalten bei Übertragungsfehlern untersucht werden.

Das vorgestellte Streaming-System für hochaufgelöste Panoramen stellt unseren ersten Schritt dar zu effizienten Streaming-Systemen für komplexe photorealistische 3D-Umgebungen, wie sie z.B. mit Image-Based-Rendering Technologien erzeugt werden.

Referenzen

[1] M. Levoy and P. Hanrahan, “Light Field Rendering”, Proc. ACM SIGGRAPH, pp. 31-42, August 1996.

[2] H.Y. Shum and L.W. He, “Rendering with Concentric Mosaics”, Proc. ACM SIGGRAPH, pp. 299-306, August 1999.

[3] E.H. Adelson and J. Bergen, “The plenoptic function and the elements of early vision”, In Computational Models of Visual Processing, pp. 3-20, MIT Press, Cambridge, MA, 1991.

[4] S.E. Chen, “QuickTime VR – An Image-Based Approach to Virtual Environment Navigation”, Proc. ACM SIGGRAPH, pp. 29-38, August 1995.

[5] A. Smolic and J.-R. Ohm, “Robust Global Motion Estimation Using a Simplified M-Estimator Approach”, Proc. ICIP2000, IEEE International Conference on Image Processing, Vancouver, Canada, September 2000.

[6] A. Smolic, “Robust Generation of 360° Panoramic Views from Consumer Video Sequences”, to appear Proc. VIPromCom2002, IEEE

International Symposium on Video/Image Processing and Multimedia Communications, Zadar, Croatia, June 16.-19. 2002.

[7] A. Smolic and T. Wiegand, “High-Resolution Video Mosaicing”, Proc. ICIP2001, IEEE International Conference on Image Processing, Thessaloniki, Greece, October 2001.

[8] A. Smolic, Y. Guo, J. Guether and T. Selinger, “Demonstration of Streaming of MPEG-4 3-D Scenes with Live Video”, ISO/IEC JTC1/SC29/WG11, MPEG01/M7811, Pattaya, Thailand, December 2001.

[9] A. Skodras, C. Christopoulos, T. Ebrahimi, “The JPEG 2000 Still Image Compression Standard”, IEEE Signal Processing Magazine, pp. 36-58, September 2001.

[10] Thomas Pintaric, Ulrich Neumann, Albert Rizzo, “Immersive Panoramic Video”, Proceedings of the 8th ACM International Conference on Multimedia, pp. 493.494, October 2000.