Die Form der bewegten Welt einfangen

Magnor, Marcus

Max-Planck-Institut für Informatik

Selbständige Nachwuchsgruppe - Graphics - Optics - Vision
Forschungsgebiet: Informatik/Mathematik/Komplexe Systeme

Korrespondierender Autor: Magnor, Marcus
E-Mail: magnor@mpi-sb.mpg.de

Zusammenfassung

Die dynamische dreidimensionale Struktur einer belebten Szene aus nur wenigen Videokameraaufnahmen zu rekonstruieren ist ein herausforderndes und spannendes neues interdisziplinäres Forschungsgebiet der Graphischen Datenverarbeitung und der Bildverarbeitung. Neue Anwendungen wie interaktives 3D-Fernsehen und Video Surround treiben das Forschungsinteresse. Dieser Beitrag gibt einen Überblick über die modernsten Verfahren, die Form unserer bewegten Welt optisch einzufangen.

Abstract

Acquiring on-line the evolving shape of a dynamic scene from a handful of video streams may be considered one of the most challenging, but at the same time also most auspicious tasks in contemporary computer graphics and computer vision research. The anticipation of revolutionary new applications such as interactive 3D television broadcasts and free-viewpoint video motivates the ongoing work. This overview aims at giving a state-of-progress report on this lively research endeavour.

Ob Fotos des vergangenen Weihnachtsfestes oder Videofilme der letzten Urlaubsreise: Die Konservierung des zweidimensionalen visuellen Eindrucks sowohl eines Augenblicks als auch bewegter Abläufe stellt heute eine technische Alltäglichkeit dar. Auch für die Vermessung der dreidimensionalen Oberfläche statischer Objekte stehen uns mit Laserscannern und anderen Techniken erprobte Verfahren zur Verfügung. Wie aber lässt sich die sich ständig verändernde dreidimensionale Form dynamischer natürlicher Szenen erfassen ? Dieser Herausforderung stellt sich die Arbeitsgruppe „Graphics-Optics-Vision" am MPI für Informatik.

Mit vertretbarem Aufwand ist diese Aufgabe nur mit passiven optischen Aufnahmeverfahren lösbar. Konventionelle Videokameras bieten eine preisgünstige Möglichkeit, goße Mengen zweidimensionaler visueller Information über unsere Umwelt zeitkritisch und computergerecht digital aufzunehmen. Mit acht Videokameras, die in unserem Studio rund um einen Bühnenbereich herum aufgestellt sind, können wir Handlungen aus verschiedenen Blickpositionen synchronisiert aufnehmen (Abb. 1). Dank vorheriger Kalibrierung der Kameras kennen wir die Positionen der Kameras und können die Videobilder miteinander in Beziehung bringen.

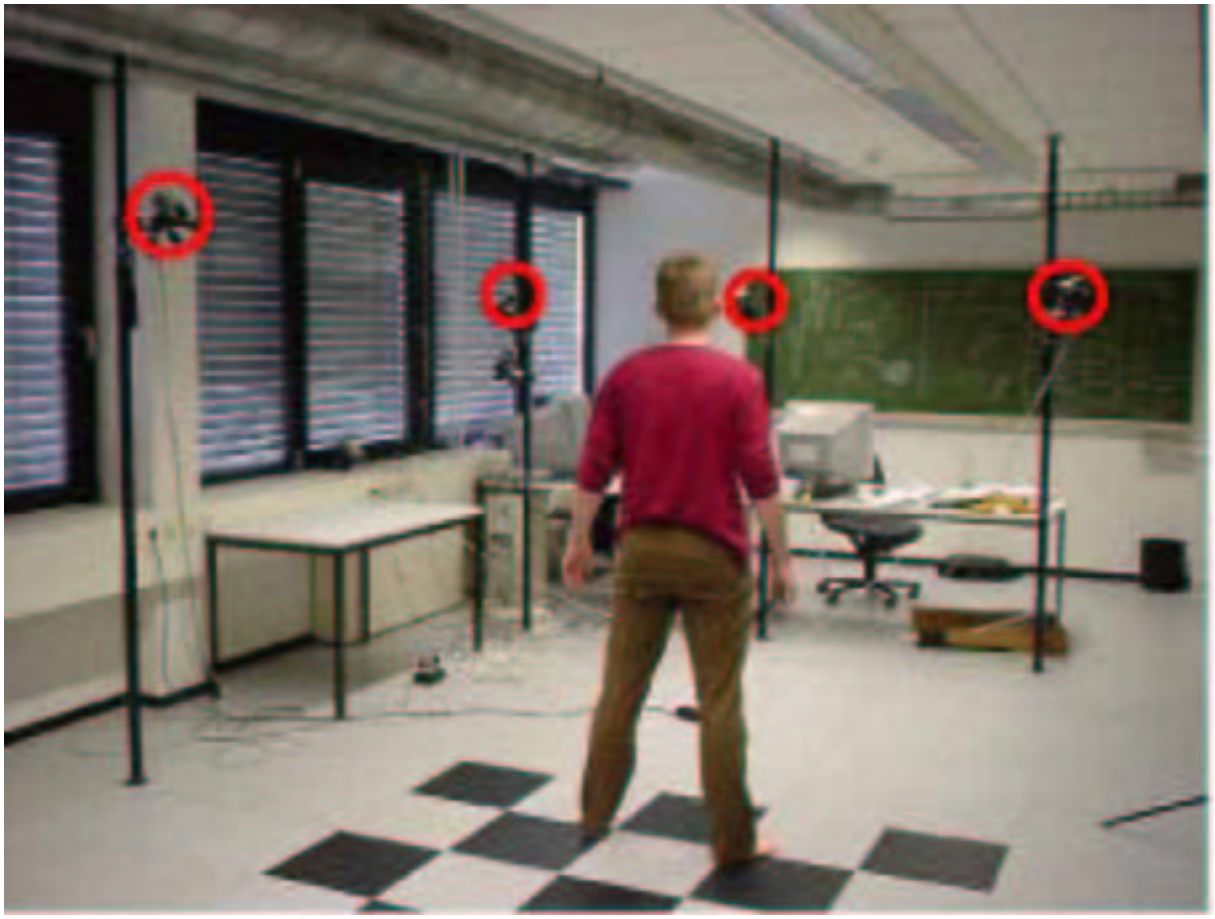

Abb. 1: *In unserem Videostudio können wir einen Bühnenbereich von acht Seiten synchron aufnehmen.*

Um aus den synchron aufgenommenen zweidimensionalen Videobildern die dynamische dreidimensionale Struktur einer Szene zu rekonstruieren, müssen Kenntnisse aus mehreren Bereichen der Informatik interdisziplinär zusammenwirken. So kennen wir aus der Bildverarbeitung Verfahren, wie aus Bildern realer Szenen dreidimensionale Strukturinformationen rekonstruiert werden können, während wir aus der Telekommunikation und Videocodierung wissen, wie mithilfe von Bildfolgen der zeitliche Ablauf analysiert wird, und die Computergrafik stellt uns Techniken zur visuellen Modellierung und realitätsgetreuen Darstellung zur Verfügung.

Durch die vorherige Aufnahme des statischen Hintergrundes können wir den Vordergrund der Szenenhandlung aus den Videobildern extrahieren. Die bekannten Kamerapositionen und Aufnahmeparameter ermöglichen es uns, diese Silhouettenbilder in den Raum zurück zu projezieren und die dreidimensionale Schnittmenge zu bestimmen. Auf diese Weise wird die optische Hülle des Vordergrundobjekts rekonstruiert (**Abb. 2**). Indem wir anschließend die Videobilder auf die optische Hülle projizieren, können wir den sich bewegenden Vordergrund aus beliebiger Ansicht wieder darstellen.

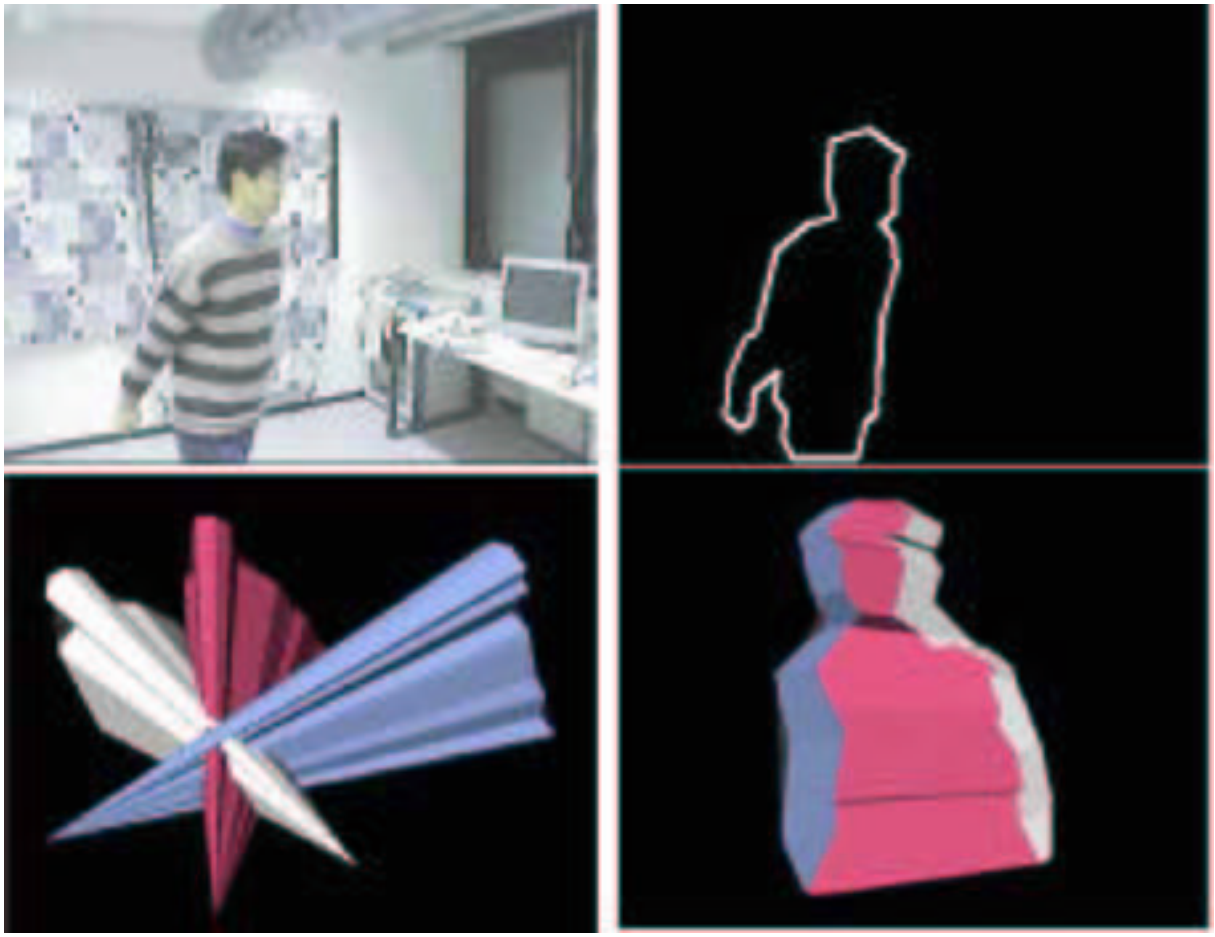

Abb. 2: *Die optische Hülle ist die Schnittmenge der Objektsilhouetten aus verschiedenen Blickrichtungen.*

Die Verarbeitung ist dabei so schnell, dass die Schauspieler auf unserer Bühne live aus jeder beliebigen Perspektive betrachtet werden können [1].

Da die optische Hülle die dreidimensionale Schnittmenge der Vordergrundsilhouetten ist, können auf diese Weise jedoch keine konkaven Objektregionen rekonstruiert werden. Außerdem haben wir nur Bilder aus acht verschiedenen Kamerapositionen zur Verfügung, sodass die rekonstruierte optische Hülle scharfe Kanten hat, die das Vordergrundobjekt nicht hundertprozentig richtig beschreiben. Durch den Vergleich der Farbwerte pro Bildelement zwischen verschiedenen Silhouettenbildern können wir jedoch zusätzlich die lokale Tiefe der Szene schätzen. Damit lässt sich die Qualität der rekonstruierten Objektgeometrie und der anschließenden Darstellung erheblich verbessern [2]. Wissen wir, was für Vordergrundobjekte sich in einer Szene bewegen, können wir mithilfe von Geometriemodellen dieses *A-priori*-Wissen dazu nutzen, die Anzahl der Freiheitsgrade während der Geometrie- und Bewegungsschätzung erheblich zu reduzieren. Zum Beispiel können wir ein Geometriemodell des menschlichen Körpers automatisch an die Silhouettenbilder anpassen, wenn wir wissen, dass unsere Szene einen Schauspieler zeigt. Die Silhouetten des Schaupielers werden dazu mit den Silhouetten des Modells verglichen und die Gelenkparameter des Modells so lange optimiert, bis die Silhouetten des Schauspielers und des Modells in allen Ansichten bestmöglich übereinander liegen

(**Abb. 3**). Dieses sehr robuste Verfahren zur Analyse menschlicher Bewegung liefert zusätzlich eine qualitativ hochwertige Objektoberfläche, die, mithilfe der Videoströme texturiert, fotorealistische Darstellungsergebnisse liefert [3].

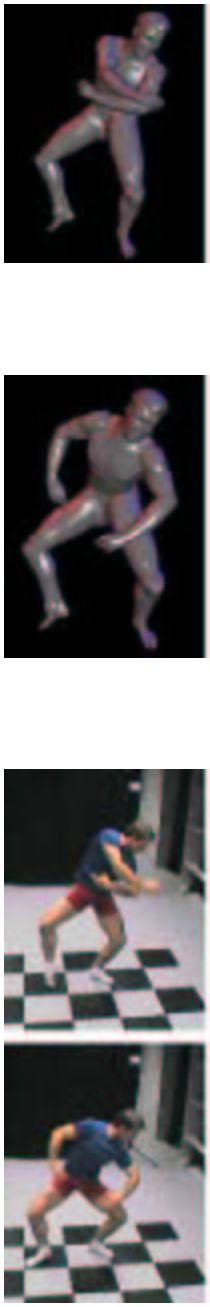

Abb. 3a - f: *Ein menschliches Geometriemodell wird fortlaufend an die Videobilder des Schauspielers angepasst, um seine Bewegungen zu analysieren.*

Stehen viele Videokameras recht eng zusammen, so sind sich die Bilder benachbarter Kameras sehr ähnlich, und nur wenige Szenenregionen sind verdeckt. In Zusammenarbeit mit der Universität Stanford, an der eine Videokameramatrix aus 100 Kameras gebaut wurde, entwickeln wir Algorithmen, die die große Ähnlichkeit dieser Videoströme nutzen, um hoch aufgelöste Tiefenkarten der Szene sehr genau und robust zu rekonstruieren (**Abb. 4**). Dabei wird der bekannte Hintergrund der Szene und die zeitliche Entwicklung der Bildfolgen ausgenutzt, um Fehlinterpretationen lokal ähnlicher Farbwerte auszuschließen [4]. Während mit der Kameramatrix nicht der gesamte Umfang

einer Szene abgedeckt werden kann, kann das Geschehen mithilfe unserer hochgenauen Tiefenkarten innerhalb des abgedeckten Raumwinkels aus beliebiger Blickrichtung in Echtzeit so gut dargestellt werden, dass die Bildqualität nicht von den tatsächlich aufgenommenen Videosequenzen zu unterscheiden ist [5].

Abb. 4a, b: *Der bekannter Bildhintergrund und die zeitliche Kohärenz des Szenenablaufs werden ausgenutzt, um genaue Tiefeninformation pro Videobildpixel zu erhalten.*

Mit unseren Forschungsarbeiten entwickeln wir die Tätigkeit des Betrachtens konservierter visueller Eindrücke von einem bislang rein passiven Vorgang weiter zu einem interaktiven Erlebnis: Anstatt sich einen Szenenablauf nur vorspielen zu lassen, steht der Betrachter nun mittendrin und ist Teil des Geschehens. Die Fußballspielübertragung kann plötzlich aus selbstgewählter Perspektive erlebt werden, aus Sicht des Stürmers oder des Schiedsrichters, des Torwarts oder auch des Balls. Alle zusätzlichen Informationen, die dazu nötig sind, liefert die raum-zeitliche Analyse der dynamischen Szenenstruktur. So überwinden wir die Kluft zwischen realer und virtueller Realität und erzeugen neue Bildeindrücke der realen, bewegten Welt.

Den aktuellen Stand unserer Forschungsarbeiten mit vielen erläuternden Texten, Bildern und Videosequenzen können Sie auf unserer Webseite http://www.mpi-sb.mpg.de/gov finden.

Literatur

[1] M. Li, M. Magnor, and H.-P. Seidel. Hardware-accelerated visual hull reconstruction and rendering. Proc. Graphics Interface (GI'03), Halifax, Canada, June 2003.

[2] M. Li, H. Schirmacher, M. Magnor, and H.-P. Seidel. Combining stereo and visual hull

information for on-line reconstruction and rendering of dynamic scenes. Proc. IEEE International Workshop on Multimedia and Signal Processing (MMSP'02), 2002.

[3] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. ACM Trans. on Computer Graphics (SIGGRAPH'03), San Diego, USA, July 2003.

[4] B. Goldlücke and M. Magnor. Joint 3-D reconstruction and background separation in multiple views using graph cuts. Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'03), Madison, USA, June 2003.

[5] B. Goldlücke, M. Magnor, and B. Wilburn. Hardware-accelerated dynamic light field rendering. Proc. Vision, Modeling, and Visualization (VMV-2002), Erlangen, Germany, pages 455–462, November 2002.