Hand-Tracking basierend auf 3D-Merkmalen

Hand-Tracking allgemein:

Die menschliche Hand als direktes Interaktionswerkzeug in virtuellen Umgebungen zu verwenden ist bereits seit Anbeginn dieses Forschungsbereichs von besonderem Interesse. Dies lässt sich durch eine der Hauptanforderungen von Virtuellen/Erweiterten Realitäten (VR/AR) erklären: Die Benutzerinteraktion soll möglichst alltägliche und damit sehr vertraute Handlungen widerspiegeln. Darüber hinaus sollen dem Benutzer Erfahrungen und Erlebnisse ermöglicht werden, die im normalen Leben unmöglich sind (z.B. Fliegen). Das tägliche Zusammenspiel mit unserer Umgebung beinhaltet nahezu unabdingbar die Verwendung unserer Hand als „Interaktionswerkzeug“, weshalb diese auch als virtuelles Gegenstück verfügbar sein sollte. Um VR/AR-Anwendungen durch bloße Handbewegungen steuern zu können, muss die Hand des Benutzers verfolgt werden; ein so genanntes Hand-Tracking muss durchgeführt werden. In diesem Kontext sind so genannte markerlose Verfahren besonders wünschenswert, damit die Interaktion unmittelbar beginnen kann und keine ungewohnten Faktoren hinzukommen (um beispielsweise Objekte greifen zu können muss wie im reellen Leben kein Zusatzinstrument getragen werden, wie z.B. ein Datenhandschuh).

Das ultimative Ziel wäre ein Hand-Tracking das die Verfolgung aller 27 Freiheitsgrade (degrees of freedom, DOF) einer Hand (globale Lage und Gelenkwinkel) erlaubt und damit alle denkbaren Interaktionsformen ermöglicht. Leider sind solche bisherigen Methoden noch nicht praxistauglich, da wichtige Anforderungen wie Echtzeitbefähigung oder automatische Initialisierung nicht erfüllt werden können. Daher beschränken sich viele andere Verfahren auf die Verfolgung einer Teilmenge von Freiheitsgraden der Hand; lediglich die globale Lage und mehrere starre Gesten werden erkannt. Auch wenn dadurch die zuvor erwähnten Probleme gelöst werden konnten, war es bislang nicht möglich die komplette räumliche Lage der Hand (Position und Orientierung) ohne Einschränkungen zu erfassen. Unser kürzlich entwickeltes Verfahren schaffte es erstmalig, alle sechs Freiheitsgrade der Lage der Hand in vier verschiedenen Gesten zu erfassen, ohne dabei solchen Einschränkungen zu unterliegen. Unser Verfahren arbeitet in Echtzeit (mindestens 25 Bilder pro Sekunde) mithilfe von drei oder mehr handelsüblichen Videokameras. Weiterhin schafft eine vollautomatische Initialisierung kombiniert mit einer stabilen Verlustdetektion (z.B. wenn sich die Hand nicht mehr im Sichtbereich befindet) alle Voraussetzungen für die Einsetzbarkeit in kommerziellen Anwendungen. Um dies zu verdeutlichen, sind bereits zwei verschiedene Applikationen von vielen verschiedenen Personen getestet worden, ein virtueller Überflug (siehe Abb. 1) und ein virtueller Zusammenbau (siehe Abb. 2). Dabei stellte sich heraus, dass die Handhabung trotz ungewohnt hohem Freiheitsgrad in der Interaktion sehr einfach und schnell zu erlernen war.

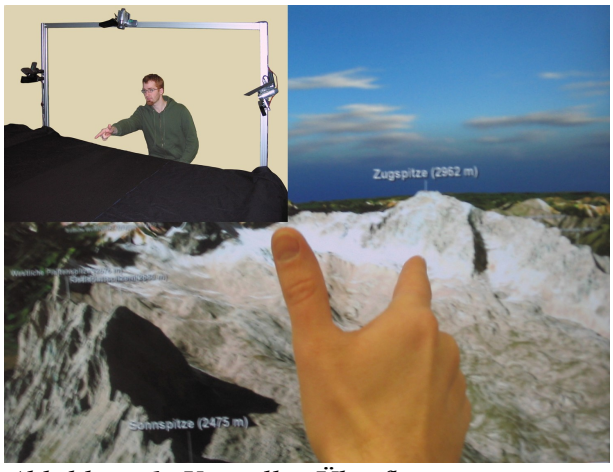

Abbildung 1: Virtueller Überflug

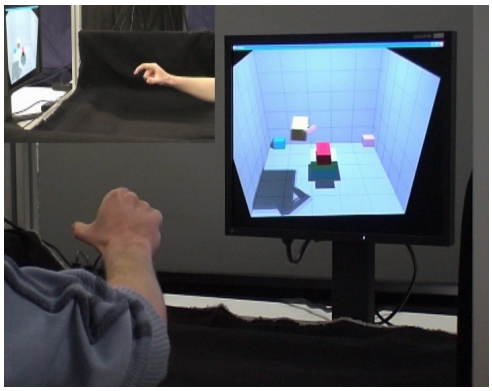

Abbildung 2: Virtueller Zusammenbau

Das Verfahren im Überblick:

Die hier vorgestellte Methode zur Verfolgung der räumlichen Lage (Position und Orientierung) der menschlichen Hand in mehreren verschiedenen Gesten mittels Techniken des maschinellen Sehens (Computer Vision) ermöglicht eine sehr natürliche und effiziente Interaktion zwischen Mensch und Maschine, da folgende Eigenschaften gewährleistet sind:

1.) Es ist ein markerloses Verfahren; der Benutzer benötigt lediglich seine bloße Hand.
2.) Die Initialisierung ist vollautomatisch. Sobald der Benutzer seine Hand in den Arbeitsbereich bewegt, startet unmittelbar die Verfolgung der Hand. Dabei ist keine spezielle Lage oder Geste der Hand erforderlich.
3.) Die Berechnung erfolgt in Echtzeit (Real-Time), so dass das Verfahren für direkte Interaktionen einsetzbar ist (mehr als 25 Updates pro Sekunde).
4.) Bei einem Benutzerwechsel sind keine Einstellungsänderungen erforderlich.

Man kategorisiert Hand-Trackingverfahren vor allem durch die Anzahl der verfolgten Freiheitsgrade. In unserem Verfahren werden sechs kontinuierliche Freiheitsgrade verfolgt, die sich aus der 3D-Position (x, y und z) der Hand und ihrer Orientierung (Drehung um x-, y- und z-Achse) im dreidimensionalen Raum zusammensetzen. In der gegenwärtigen Realisierung werden vier starre Gesten der Hand unterschieden und erkannt (siehe Abb. 4).

Drei oder mehr (für erhöhte Stabilität/Genauigkeit) kalibrierte Kameras beobachten in spezieller Anordnung (Winkel von 60° bei drei Kameras, siehe Abb. 3 oben links) die Hand aus verschiedenen Richtungen. Der von allen Kameras gleichzeitig einsehbare Bereich stellt das Arbeitsvolumen dar, in welchem die Hand verfolgt wird. Dabei gilt: Je größer das Arbeitsvolumen (durch größeren Sichtwinkel oder Distanz der Kameras), desto ungenauer ist das Tracking (bedingt durch eine niedrigere Auflösung der Hand in den Kamerabildern). Ausreichende Genauigkeiten können jedoch noch bei einer Größe von zwei Kubikmetern unter Verwendung von Kameras mit Auflösungen von 720x576 Bildpunkten erreicht werden.

Um nun die Lage und Geste der Hand zu errechnen, wird zunächst eine hinreichend genaue 3D-Rekonstruktion der Hand aus den Kamerabildern bestimmt und die zweidimensional erfasste Information in eine konsistente dreidimensionale Darstellung gebracht. Dazu werden in der aktuellen Realisierung die Bilder aller drei Kameras synchron ausgelesen und jeweils in Handregion und Hintergrund unterteilt, also segmentiert. Damit diese Trennung durch einfache Hintergrund Subtraktion (Background Subtraction) erreicht werden kann, wird der Sichtbereich der Kameras mit Infrarotlicht beleuchtet, ein nicht Infrarot reflektierendes Tuch im Hintergrund ausgelegt und die mit Infrarotfiltern ausgestatteten Kameras im Nachtmodus betrieben (siehe Abb. 1 links oben). Sobald alle Bilder segmentiert sind, werden die Handregionen ausgehend vom Blickpunkt der jeweiligen Kamera durch den dreidimensionalen Raum projiziert so dass sich im Schnitt der drei Projektionen eine grobe 3D-Rekonstruktion der Hand ergibt (siehe Abb. 3 oben). Diese Methode heißt

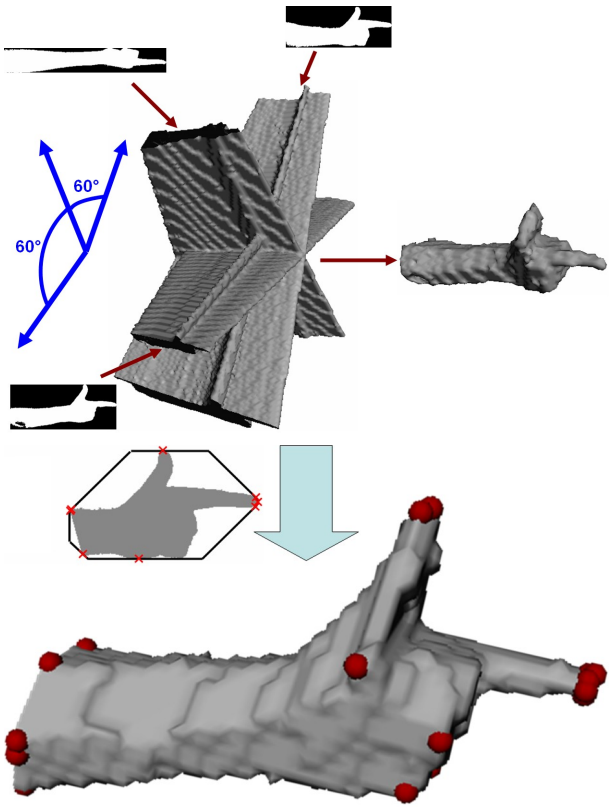

Abbildung 3: Rekonstruktion der Visual Hull (oben) und hervorstehende Merkmale (unten)

Rekonstruktion der Visual Hull oder *Shape from Silhouettes Technik*.
In der groben 3D-Rekonstruktion der Hand wird nun nach besonderen Merkmalen (Features) gesucht. Um die Menge der Kandidaten einzuschränken, wird nur nach hervorstehenden Merkmalspunkten gesucht, wie den Fingerspitzen, die auf einem k-DOP, einer Approximation der konvexen Hülle der Hand, liegen (siehe Abb. 3 unten). Ein k-DOP (Diskretes orientiertes Polytop, discrete oriented polytope) ist ein Hüllvolumen (bounding volume), welches konstruiert wird, indem k wohlorientierte Ebenen aus dem Unendlichen bewegt werden bis sie die 3D-Rekonstruktion berühren. Das k-DOP ist dann dasjenige konvexe Polytop, welches aus dem Schnitt der Halbräume resultiert, die durch diese k Ebenen begrenzt werden (für ein Beispiel im 2D siehe Abb. 3 Mitte). Für jede dieser Ebenen gibt es also einen zur 3D-Rekonstruktion gehörenden Voxel (Volumetrischer Pixel), der die Ebene berührt und damit ihre Position beschreibt. In der aktuellen Realisierung des Verfahrens wurde ein 26-DOP verwendet, so dass es 26 Ebenen bzw. Orientierungen gibt und damit 26 Voxel bestimmt werden. Diese 26 Voxel bilden unsere Kandidatenmenge für die Extraktion der Fingerspitzenmerkmale und werden nun klassifiziert, indem ihre lokalen Umgebungen analysiert werden. Für unser Verfahren haben wir uns entschieden nur eine sehr einfache Analyse durchzuführen, so dass lediglich der Abstand zum lokalen Masseschwerpunkt zur Charakterisierung verwendet wird. Ist der Abstand sehr groß, so befindet sich der Voxel bzw. das Merkmal auf einem sehr hervorstehenden Teil der 3D-Rekonstruktion, und damit wahrscheinlich auf einer der gewünschten Fingerspitzen. Manchmal kann es vorkommen, dass sich auch hervorstehende Merkmale bilden, wenn der Arm vom Kamerabild schräg abgeschnitten wird. Diese Merkmale werden im Nachhinein aussortiert, indem die 3D-Positionen der Voxel in die 2D-Kamerabilder projiziert werden und geprüft wird, ob die projizierte Position am Rand des Bildes liegt. Liegt sie am Rand, so wird das Merkmal verworfen. Nachdem die zwei hervorstehendsten Merkmale als Kandidaten für die zwei Fingerspitzen erkannt wurden, wird durch weitere Methoden zwischen vorderem Finger und Daumen unterschieden sowie der Handmittelpunkt berechnet. Durch diese drei Punkte ist die Lage der Hand hinreichend bestimmt (siehe Abb. 4 oben). Außerdem werden weitere Analysen, wie z.B. Hauptkomponentenanalyse (Principal Component Analysis, PCA), der lokalen Umgebungen der Fingerspitzen durchgeführt um die verschiedenen Gesten zu klassifizieren.

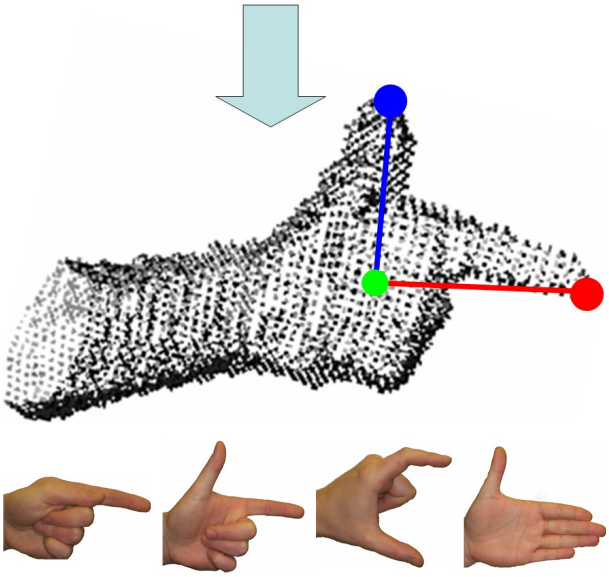

Abbildung 4: Handlage (oben) und Handgesten (unten)

Weitere Informationen und Kontakt:

Universität Bonn – Institut für Informatik II – Arbeitsgruppe Computer Graphik
http://cg.cs.uni-bonn.de/projects/markerless-hand-tracking

Relevante Literatur:

M. Schlattmann, F. Kahlesz, R. Sarlette, R. Klein 2007. Markerless 4 gestures 6 DOF real-time visual tracking of the human hand with automatic initialization. In Computer Graphics Forum 26(3):467-476. Eurographics Association.

M. Schlattmann, R. Klein 2007. Simultaneous 4 gestures 6 DOF real-time two-hand tracking without any markers. In proceedings of ACM Symposium on Virtual Reality Software and Technology.