

# **Monokulare Modellbasierte Posturschätzung des Menschlichen Oberkörpers**

Joachim SCHMIDT

## **Zusammenfassung**

In diesem Papier wird ein modellbasiertes Verfahren zur 3D Posturschätzung des menschlichen Oberkörpers vorgestellt. Dieses Verfahren ist auch in monokularen Bildern ohne explizite Distanzmessungen in der Lage, die Position eines Menschen und einzelner Körperteile im 3D Raum zu bestimmen. Die Entfernung ergibt sich dabei implizit durch die Hinzunahme von Modellwissen. Das vorgestellte System verwendet zur Verfolgung einer Postur über die Zeit einen Kernel-Partikelfilter. Dieses probabilistische Suchverfahren nutzt verschiedene Farb- und Intensitätsmerkmale zur Bewertung der Übereinstimmung zwischen Modell und Person im Bild. Die Evaluation erfolgt anhand einer Beispielsequenz, die einen Menschen zeigt, der verschiedene Montagehandlungen an einem Motorblock durchführt. Die verfolgten Bewegungen werden mit Ground Truth Daten verglichen, die aus photogrammetrischen Marken unter Zuhilfenahme einer trinokularen Kamera ermittelt wurden.

## **1 Szenario und Aufgabenstellung**

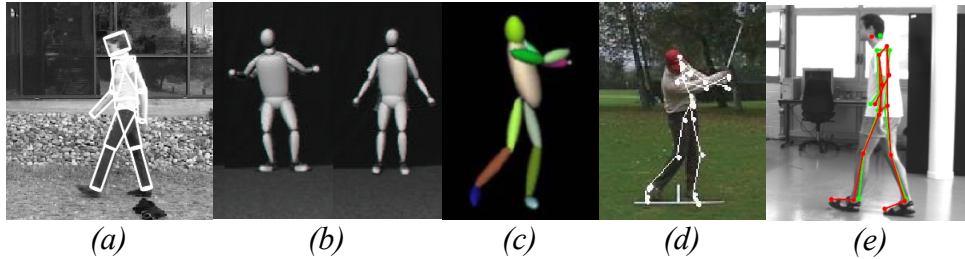
Ziel dieser Arbeit ist die Analyse der Bewegungen einer Person, die verschiedene Montagehandlungen an einem Motorblock durchführt, zum Beispiel das Festziehen von Schrauben mit einem Schraubenschlüssel. Die Aufgabe besteht in der Verfolgung der 3D-Postur des Oberkörpers der Person, insbesondere des rechten Armes und der rechten Hand.

Die verwendete Videosequenz wurde mit einer multiokularen Farbkamera aufgenommen. Zur Analyse der Bewegungen wird jedoch nur eine einzelne Kamera verwendet, woraus sich eine besondere Herausforderung ergibt: Die Entfernung eines Objektes oder einer Person kann nicht mittels Triangulation bestimmt werden, was der übliche Ansatz bei multiokularen Verfahren, z.B. bei der Tiefenbestimmung über Punktkorrespondenzen ist. Vielmehr stellt diese Arbeit ein Verfahren vor, welches auch ohne explizite Distanzmessungen in der Lage ist, die Position eines Menschen und einzelner Körperteile im 3D Raum zu bestimmen. Die Entfernung ergibt sich dabei implizit durch die Hinzunahme von Modellwissen.

## **2 Verwandte Arbeiten zur Erkennung von Körperposturen**

Die Grundidee der modellbasierten monokularen Posturschätzung besteht darin, ein Modell der zu beobachtenden Person zu nutzen, welches für ein aufgenommenes Bild Rückschlüsse auf die Postur der Person zulässt. Das Modell bildet dabei sowohl die Bewegungsmöglichkeiten der Person als auch ihr Erscheinungsbild nach. Indem nun das Modell in eine Po-

sition gebracht wird, welche mit dem aktuellen Bildeindruck möglichst exakt übereinstimmt, kann die Postur des Menschen ermittelt werden.



**Abb. 1:** Erkennungsergebnisse verwandter Verfahren zur Verfolgung von Körperposturen aus monokularen Bildfolgen. Bilder entnommen aus: (a) SIDENBLADH (2001), (b) SMINCHISESCU (2001), (c-d) URTASON (2005), (e) LU (2008)

Ein solcher Ansatz zur Verfolgung eines 3D Modells des menschlichen Körpers auf der Basis eines Partikelfilters wurde von SIDENBLADH (2001) präsentiert, siehe Abb. 1(a). SMINCHISESCU (2001) verwendet ein detaillierteres Modell und ein komplexeres Verfahren zur Exploration des hochdimensionalen Suchraumes, vgl. Abb. 1(b). Falls Wissen über die beobachtete Handlung vorhanden ist, so kann dies gewinnbringend genutzt werden, um die Erkennungsleistung zu verbessern. Dies zeigen u.a. die Arbeiten von URTASON (2005), siehe Abb. 1(c-d), der den Abschlag eines Golfspielers untersucht und auch LU (2008), siehe Abb. 1(e), welcher eine Möglichkeit zur Dimensionsreduktion des Rekonstruktionsproblems aufzeigt, indem die Schätzung zugunsten bekannter Posen und Bewegungen beeinflusst wird.

### 3 Modellbasierte Monokulare Posturschätzung

Dieses Papier stellt einen Ansatz zur Schätzung der Körperpostur eines Menschen aus einer monokularen Bildfolge vor. Der Ansatz basiert auf der Verwendung eines beweglichen Modells zur Repräsentation des Erscheinungsbildes einer Person und dem Abgleich einzelner Stellungen dieses Modells - auch Posturen genannt - mit Hilfe von verschiedenen Bildmerkmalen. Die effiziente zielgerichtete Suche nach der besten Postur wird mit einem Kernel-Partikelfilter erreicht, vgl. SCHMIDT (2006), der mit Hilfe von Bewegungsmodellen die wahrscheinliche nächste Position des Modells vorhersagt. Die verschiedenen Schritte des Verfahrens zur modellbasierten monokularen Posturschätzung werden im Folgenden erläutert.

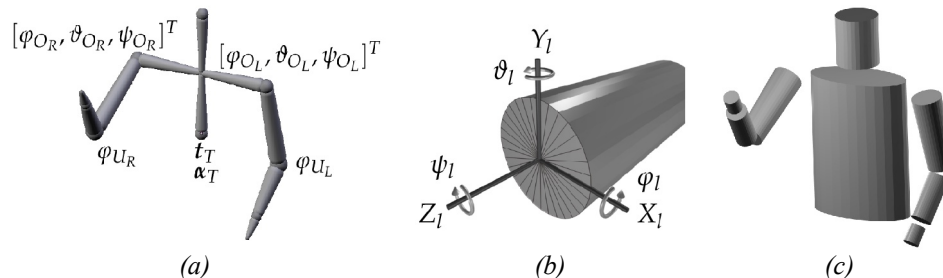
#### 3.1 Aufbau des Algorithmus

Ausgehend von einer initialen Modellpostur oder dem letzten Erkennungsergebnis wird durch den Kernel-Partikelfilter eine neue Menge von Körperposturhypothesen erzeugt. Dies kann unter Zuhilfenahme eines Bewegungsmodells geschehen, welches die nächste Postur anhand der bereits beobachteten Bewegung vorhersagt. Alle Hypothesen werden mittels

des Meanshift Verfahrens optimiert. Dazu wird zunächst für jede Hypothese die Wahrscheinlichkeit  $p(y_t|x_t)$  der Übereinstimmung zwischen Bildeindruck und gegebener Modellpostur anhand verschiedener Bildmerkmale berechnet. Ist die gesamte Menge der Körperposturen bewertet, so werden sie jeweils in Bereiche mit hoher Übereinstimmung verschoben, um diese Regionen des Parameterraumes in den nachfolgenden Iterationen noch genauer abtasten zu können. Dieser Prozess des Bewertens und Verschiebens wird so lange fortgesetzt, bis keine Verbesserung bei der Bewertung mehr erkennbar ist oder eine Maximale Anzahl an Iterationen erreicht wurde. Aus der Menge aller verbesserten Modellposturen kann nun eine einzelne Postur ermittelt werden, die die Stellung des Oberkörpers der Person für das aktuelle Bild am besten beschreibt. Diese Postur wird als Ergebnis für den aktuellen Zeitpunkt ausgegeben und das gesamte Verfahren für das nächste Bild fortgesetzt.

### 3.2 Körpermodell

Durch die Verwendung eines 3D Modells ist man in der Lage, sogar aus einem 2D Bild Entfernungen zu bestimmen. Die Position der Person im Raum und auch die Entfernung einzelner Körperteile ergibt sich dabei nicht direkt durch Messung, sondern implizit aus mehreren Faktoren, wie der Größe im Bild in Kombination mit der Körperhaltung. So kann z.B. bei einem angewinkelten Arm mit der Hand vor dem Bauch die Hand nur in einer gewissen Entfernung vom Körper sein, was man an der Durchstreckung des Ellenbogengelenkes erkennen kann. Solche Informationen sind auch im 2D Bild gut erkennbar, jedoch lassen sie sich nur unter Zuhilfenahme eines 3D Körpermodells auswerten. Ebenso können unmögliche Stellungen - z.B. Hand im Bauch - und sehr unwahrscheinliche Stellungen - z.B. Hand hinter dem Rücken - ausgeschlossen werden, wodurch der Suchraum eingeschränkt wird. Zusätzlich können eventuelle Mehrdeutigkeiten mit Hilfe des Modells aufgelöst werden.



**Abb. 2:** Bewegliches Körpermodell mit 14 Freiheitsgraden. (a) Kinematisches Modell, (b) Körperteil repräsentiert als Zylinder, (c) resultierendes Modell.

Die Voraussetzung dafür ist ein Modell des zu beobachtenden Menschen, welches sowohl die Körpermaße möglichst exakt nachbildet, als auch gleichzeitig möglichst gut generalisiert, um unvorhersagbare Änderungen des Bildeindruckes, z.B. durch wechselnde Beleuchtungsbedingungen oder Falten in der Kleidung zu tolerieren. Der Aufbau des hier verwendeten Modells ist in Abb. 2 dargestellt. Eine Postur des Modells lässt sich vollständig

durch den Parametervektor  $\mathbf{x}_t$  beschreiben der sich aus der Position des Modells im Raum und der Stellung der einzelnen Gelenke zusammensetzt:

$$\mathbf{x}_t = [\varphi_T, \vartheta_T, \psi_T, x_T, y_T, z_T, \varphi_{U_R}, \vartheta_{U_R}, \psi_{U_R}, \varphi_{L_R}, \varphi_{U_L}, \vartheta_{U_L}, \psi_{U_L}, \varphi_{L_L}]^T \quad (1)$$

### 3.3 Bewertung einer Postur durch Fusion von Bildmerkmalen

Mittels verschiedener Intensitäts- und Farbmerkmale kann das aktuelle Bild mit der Postur des Modells verglichen werden. Die Bildmerkmale bilden die Schnittstelle zwischen der abstrakten Repräsentation des Körpermodells und dem tatsächlichen Bildeindruck. Sie geben somit eine Art Erwartungshaltung vor, die beschreibt, welcher Bildeindruck bei der aktuellen Postur des Modells erwartet wird.

Die Bildmerkmale - im Folgenden auch Filter genannt - werden dazu auf verschiedenen Bereichen des Körpermodells berechnet, je nachdem ob sie eine erwartete Kante oder Fläche modellieren. Nicht jedes Merkmal wird daher für jedes einzelne Körperteil berechnet. Vielmehr kann eine geschickte Auswahl der Art und Anzahl der zu berechnenden Merkmale für ein einzelnes Körperteil entscheidend zur Robustheit der Bewertungsfunktion beitragen. Für diese Arbeit kommen unterschiedliche Merkmale zum Einsatz, deren Eigenschaften im Folgenden kurz zusammengefasst werden. Für eine detailliertere Betrachtung der Merkmale und des Algorithmus sei auf SCHMIDT (2006) verwiesen.

#### 3.3.1 Eigenschaften der Filter

Farbe ist ein sehr aussagekräftiges Merkmal, insbesondere wenn sich die Kleidung der Person deutlich von der Farbe der Umgebung unterscheidet. Das regionenbasierte Farbmittelwert-Merkmal modelliert das Aussehen der Körperteile zum Zeitpunkt  $\mathbf{t}$ , indem der Farbmittelwert  $C_t$  für mehrere Regionen  $\mathbf{b}$  auf jedem Körperteil  $\mathbf{l}$  mit einem gelernten Farbmodell  $\bar{C}_{t-1}$  verglichen wird, siehe auch Abb. 3(c).

$$f_C^{(b,l)} = \left\| C_t(z^{(b,l)}) - \bar{C}_{t-1}^{(b,l)} \right\| \quad (2)$$

Auch Hautfarbe ist ein sehr spezifisches Merkmal, da es in der Umgebung nur selten vorkommt. Somit kann eine Segmentierung des Eingabebildes nach Hautfarbe einen guten Hinweis liefern, wo sich die Hände und das Gesicht des Menschen befinden. Der hier verwendete Algorithmus zur Hautfarbensegmentierung nach FRITSCH (2002) klassifiziert jedes Pixel nach Hautfarbe oder Hintergrund mit Hilfe eines Mischverteilungs-Klassifikators im RG-Farbraum. Der verwendete RG-Farbraum hat den Vorteil, dass er helligkeitsinvariant ist, d.h. Pixel werden ausschließlich nach ihrer spezifischen Farbe segmentiert und nicht nach ihrer absoluten Helligkeit. Ähnlich zu dem vorherigen Merkmal werden auch hier mehrere Regionen  $\mathbf{b}$  pro Körperteil  $\mathbf{l}$  genutzt, um die Filterantwort zu generieren, siehe Abb. 3(c). Die Filterantwort berechnet sich aus dem Verhältnis von hautfarbenen und nicht hautfarbenen Pixeln  $\mathbf{z}_m$  innerhalb einer Region, wobei  $\psi(z_m)=1$  liefert, wenn das Pixel hautfarben ist und  $\mathbf{0}$ , falls nicht.

$$\bar{f}_S^{(l)} = \frac{1}{M_S} \sum_{m=1}^{M_S} \psi(z_m^{(b,l)}) \quad (3)$$

Nachdem die beiden vorherigen Merkmale die Oberfläche einzelner Körperteile beschreiben, wird mit dem folgenden Merkmal nach der Kante eines Körperteils gesucht. Wir erwarten, dass hier ein Wechsel zwischen Vordergrund und Hintergrund sichtbar wird, was sich üblicherweise durch einen abrupten Intensitätsunterschied bemerkbar macht. Dazu wird nicht nur das alleinige Vorhandensein einer Kante, sondern auch deren Ausrichtung berücksichtigt. Über den Gradienten  $\nabla f(x,y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right)$  wird die Kantenrichtung im Bild mit dem Winkel  $\alpha$  der Kante des Körperteils verglichen. Dies geschieht für mehrere Punkte, die äquidistant auf der Kante des Modells angeordnet sind, siehe Abb. 3(c). Die Filterantwort des Kantenmerkmals für einen Merkmalspunkt berechnet sich durch:

$$f_E^{(l)}(z^{(m)}, \alpha) = \partial_y(z^{(m)}) \cos(\alpha) - \partial_x(z^{(m)}) \sin(\alpha) \quad (4)$$

Das Profilvermerkmal wird benutzt, um längliche Strukturen eines bestimmten Durchmessers - ein sogenanntes Profil - im Bild zu finden, wie z.B. einen Arm. Da das Merkmal abhängig von der Größe des jeweiligen Körperteils im Bild ist, kann es eigentlich nur dann korrekte Ergebnisse liefern, wenn das beobachtete Körperteil den richtigen Abstand zur Kamera hat. Um bei beliebigen Entfernungen zu funktionieren, wird das Merkmal nicht auf dem Ursprungsbild angewendet, sondern auf einer Gaußpyramide. Dabei wird abhängig von der Entfernung des Modells die passende Verkleinerungsstufe der Pyramide ausgewählt, so dass das Körperteil im Bild in der richtigen Größe erscheint. Das Profilvermerkmal unterdrückt punktförmige Intensitätssprünge, die z.B. durch Kamerarauschen entstehen können, indem es parallel zum Körperteil verlaufende Gradienten positiv bewertet. Gleichzeitig werden Gradienten, die senkrecht zur Ausrichtung des Körperteils verlaufen schlecht bewertet. Dazu werden wie schon beim Kantenmerkmal mehrere Merkmalspunkte ausgewertet, die hier jedoch in der Mitte des jeweiligen Körperteils liegen, siehe Abb. 3(c).

$$f_R^{(l)}(z, \alpha) = \left| \sin(\alpha)^2 \partial_{xx}^{(\mu)}(z) + \cos(\alpha)^2 \partial_{yy}^{(\mu)}(z) - 2 \sin(\alpha) \cos(\alpha) \partial_{xy}^{(\mu)}(z) \right| - \left| \cos(\alpha)^2 \partial_{xx}^{(\mu)}(z) + \sin(\alpha)^2 \partial_{yy}^{(\mu)}(z) + 2 \sin(\alpha) \cos(\alpha) \partial_{xy}^{(\mu)}(z) \right| \quad (5)$$

### 3.3.2 Merkmalsfusion

Je nach Situation in der eine Person beobachtet wird, entstehen mitunter sehr verschiedene Bildeindrücke des gleichen Menschen. Auch die zur Berechnung herangezogenen Merkmale reagieren dann sehr unterschiedlich. Ein einzelnes Merkmal kann in einer Situation noch sehr aussagekräftig sein, bei einer anderen Ansicht aber kaum noch sinntragende Informationen liefern. Die Auswertung mehrerer Bildmerkmale und deren anschließende Fusion ist der Versuch, die Vorteile der verschiedenen Merkmale zu kombinieren. Selbst wenn einige wenige Merkmale unzureichende oder gar falsche Informationen liefern, so ist durch eine geschickte Kombination aller Merkmale trotzdem ein stabiler Vergleichswert berechenbar.

Jedes Merkmal (abgekürzt mit  $c \in \{M, S, E, R\}$ ) liefert zunächst lediglich eine Filterantwort  $f_c^{(l)}$  zurück, die mittels einer Gaußfunktion in eine Wahrscheinlichkeit überführt wird. Diese sogenannte Transferfunktion bildet mittels einer Exponentialfunktion aus dem Wertebereich des jeweiligen Merkmals in den einheitlichen Wertebereich  $[0,1]$  ab:

$$p(c, l) = \exp\left(-\frac{(\bar{f}_c^{(l)})^2}{2\sigma_c^2}\right) \quad (6)$$

Der Wert von  $p(c, l)$  gibt somit die Wahrscheinlichkeit an, dass sich das Körperteil **l** des Modells an der korrekten Position im Bild befindet. Je höher die Wahrscheinlichkeit, desto besser die Übereinstimmung. Die Standardabweichung  $\sigma_c$  wird für jedes Merkmal empirisch ermittelt.

Für die Bewertung einer Gesamtpostur werden die Einzelwahrscheinlichkeiten in einer Verbundwahrscheinlichkeit zusammengefasst:

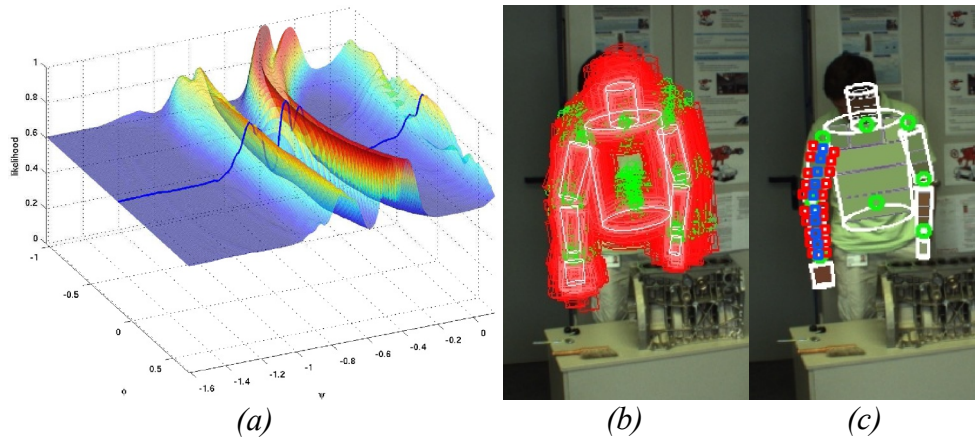
$$p(\mathbf{y}_t | \mathbf{x}_t) = \prod_{c \in \{E, R, M, S\}} \prod_{l=1}^L p(c, l)^{\frac{1}{\lambda_c N_c^{(l)}}} \quad (7)$$

Diese gibt die Wahrscheinlichkeit an, dass bei einer gegebenen Modellpostur  $\mathbf{x}_t$  die Observation  $\mathbf{y}_t$  vorliegt. Da die Anzahl der verwendeten Merkmale  $N_c^{(l)}$  für jedes Körperteil unterschiedlich sein kann, wird diese als Normierungsfaktor berücksichtigt. Der Faktor  $\lambda_c$  ist eine merkmalspezifische Gewichtungskonstante. Diese kann dazu verwendet werden, die Merkmale untereinander zu gewichten, da manche Merkmale in gewissen Situationen besser für die Schätzung einer Postur geeignet sind als andere. Dieser Parameter ermöglicht somit bei Bedarf eine Anpassung an verschiedene Beobachtungssituationen, z.B. kann bei einem stark strukturierten Hintergrund das Kantenmerkmal weniger stark gewichtet werden oder bei einer besonders hervorstechenden Kleidung das Farbmerkmal entsprechend höher.

### 3.4 Posturschätzung als Optimierungsproblem

Mit der Definition des Körpermodells und der dazugehörigen Bildmerkmale kann die Aufgabe der Bestimmung der Postur des Menschen nun zurückgeführt werden auf die Suche nach der besten Übereinstimmung von Modellzustand und Bild. Das Ergebnis dieser Suche ist eine Schätzung der Körperpostur für das aktuelle Bild. Für die gesamte Videosequenz ergibt sich somit eine Folge von Einzelposturen, die als Bewegung interpretiert werden kann.

Das vorgestellte System verwendet zur Verfolgung einer Postur über die Zeit einen Kernel-Partikelfilter, siehe SCHMIDT (2006). Der Kernel-Partikelfilter ist ein probabilistisches Suchverfahren und nutzt die in Gleichung 7 dargestellte Gesamtbewertung zur Bewertung der Übereinstimmung zwischen Modell und Person in der Szene. Die Suche nach der besten Postur geschieht über Variation des Parametervektors. Das Verfahren versucht nun das Maximum der bedingten Wahrscheinlichkeitsdichte  $p(\mathbf{x}_{t-1} | \mathbf{Y}_{t-1})$  aller Modellposturen zum Zeitpunkt  $\mathbf{t}$  zu ermitteln. Eine vollständige Suche im 14-dimensionalen Parameterraum gestaltet sich jedoch als schwierig, da dazu eine sehr große Anzahl von Hypothesen getestet werden müsste, die Berechnung der Filterantworten  $f_c^{(l)}$  aller Merkmale jedoch mit einem hohen Rechenaufwand verbunden ist. Weiterhin kann nicht sichergestellt werden, dass das globale Optimum auch der tatsächlichen Postur im Bild entspricht, da durch Mehrdeutigkeiten oft falsche lokale Optima entstehen.



**Abb. 3:** Suche nach der besten Postur. (a) Wahrscheinlichkeitsdichte des Profilmerkmals für zwei Freiheitsgrade im Parameterraum, (b) Rückprojektion der Partikelverteilung als Posturhypothesen, eingefärbt nach ihrer Wahrscheinlichkeit der Übereinstimmung, (c) beste Postur mit Regionen des Farbmittelwertmerkmals, sowie die Positionen des Kanten- (rot) und Profilmerkmals (blau).

Aus diesem Grund wird die Wahrscheinlichkeitsdichte durch eine Menge von Partikeln  $\mathbf{S}_t = \{\mathbf{s}_t^{(1)}, \dots, \mathbf{s}_t^{(N)}\}$  approximiert. Jedes Partikel  $\mathbf{s}_t^{(n)}$  entspricht dabei einer möglichen Modellpostur und erhält entsprechend Gleichung 7 eine Bewertung, auch Gewicht genannt. Die resultierende Wahrscheinlichkeitsdichte ist in Abbildung 3(a) exemplarisch für zwei Freiheitsgrade des Parameterraumes für das Profilmerkmal dargestellt. Abbildung 3(b) zeigt die Partikelverteilung als Menge von Körperposturen, die entsprechend ihrer jeweiligen Bewertung eingefärbt wurden.

Mit Hilfe der Observationen  $\mathbf{Y}_{t-1} = \{\mathbf{y}_0, \dots, \mathbf{y}_{t-1}\}$  aus den vorangegangenen Zeitschritten propagiert der Partikelfilter die bedingten Wahrscheinlichkeiten  $p(\mathbf{x}_{t-1} | \mathbf{Y}_{t-1})$  in den aktuellen Zeitschritt. Hierzu werden die Partikel  $\mathbf{S}_{t-1}$  aus dem vorherigen Zeitschritt aufgrund ihrer Gewichte  $\{w_{t-1}^{(n)}\}_{n=1}^N$  probabilistisch ausgewählt, anhand eines Bewegungsmodells in den aktuellen Zeitschritt gestreut und die Gewichte  $\{w_t^{(n)}\}_{n=1}^N$  von  $\mathbf{S}_t$  neu bestimmt.

Die hohe Dimensionalität des Problems bedingt eine sehr geringe Abdeckung des Suchraums durch die Partikel. Dadurch kann es passieren, dass gute Lösungen in der unmittelbaren Umgebung des gesuchten Maximums nicht beachtet werden. Ein Ausweg wäre die Anzahl der Partikel zu erhöhen, was jedoch aufgrund des resultierenden hohen Rechenaufwands möglichst vermieden werden sollte und weiterhin aufgrund des hochdimensionalen Merkmalsraums auch nur bedingt sinnvoll ist. Daher wendet der Kernel-Partikelfilter ein Meanshift-Verfahren zur iterativen Verdichtung der Partikelverteilung an, um die lokale Dichte um die Maxima zu erhöhen.

Der Meanshift-Algorithmus, wie von COMANICIU (2002) ist ein kernelbasiertes Verfahren, das die Dichteverteilung der Partikel mit Hilfe einer Fensterfunktion  $H_h(\cdot)$  approximiert.

Das vorgestellte System verwendet den radialsymmetrischen Epanechnikov-Kernel. Das Meanshift-Verfahren berechnet einen gewichteten Mittelwert

$$\mathbf{m}(\mathbf{s}_t^{(n)}) = \frac{\sum_{i=1}^N H_h(\mathbf{s}_t^{(n)} - \mathbf{s}_t^{(i)}) w_t^{(i)} \mathbf{s}_t^{(i)}}{\sum_{i=1}^N H_h(\mathbf{s}_t^{(n)} - \mathbf{s}_t^{(i)}) w_t^{(i)}} \quad (8)$$

für jeden Partikel  $\mathbf{s}_t^{(n)}$  und verschiebt diesen dann mittels des sogenannten Meanshift-Vektors  $\mathbf{s}_{t,neu}^{(n)} = \mathbf{m}(\mathbf{s}_t^{(n)}) - \mathbf{s}_t^{(n)}$  in einen Bereich mit höherer Dichte. Die Gewichte  $\{w_t^{(n)}\}_{n=1}^N$  müssen dazu nach jeder Iteration neu berechnet werden. Von jeder beliebigen Position im Merkmalsraum wird das nächstliegende lokale Maximum gefunden. Das Meanshift-Verfahren wird so lange fortgesetzt, bis entweder eine feste Anzahl von Iterationen erreicht ist oder keine merkliche Verschiebung mehr auftritt. Aus den resultierenden Maxima können nun Modellposturen mit einer guten Bewertung ermittelt werden. Über die Zeit ergibt sich somit eine Folge von Einzelposturen, die als Bewegung des Menschen interpretiert werden kann.

## 4 Ergebnisse

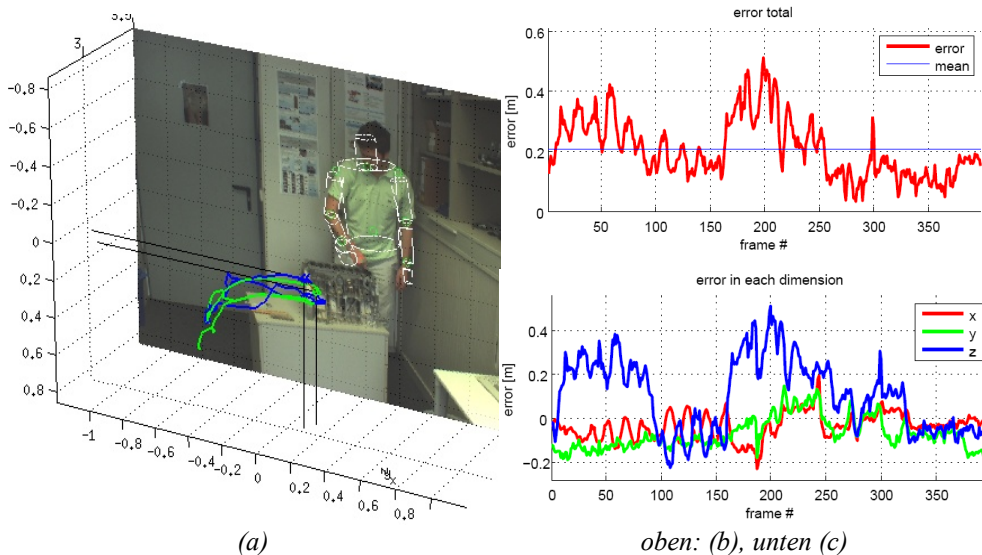
Das Verfahren zur Posturschätzung wurde für die erwähnte Videosequenz evaluiert. Daimler hat zusätzlich zu den eigentlichen Bildern 3D Markerdaten mitgeliefert, die als Ground Truth verwendet werden können. Dadurch ist ein direkter Vergleich zwischen der getrackten Modellposition und der tatsächlichen Position der Person im Bild möglich.

Für die im Folgenden vorgestellte Evaluierung wird der Marker, der auf dem Rücken der rechten Hand angebracht ist, mit der Position der rechten Hand des Körpermodells verglichen, siehe auch Abbildung 4(a). Dabei treten systematische Fehler auf. Diese entstehen durch den Versatz der zu vergleichenden Punkte, da sich der Marker auf der Hautoberfläche befindet und der Vergleichspunkt auf dem Modell an der „Fingerspitze“ der Hand. Der dadurch zu erwartende Fehler liegt bei wenigen cm. Weiterhin ist das monokulare Verfahren nicht in der Lage, die Position vom Punkten im Raum direkt zu messen, vielmehr ergibt sich diese indirekt aus der Skalierung des Körpermodells und unter Ausnutzung der bekannten Kamerageometrie. Da die Maße der zu beobachtenden Person im Vorfeld nicht bekannt waren, wurde hier ein generisches Körpermodell genutzt und dessen initiale Postur und Skalierung entsprechend des Startbildes gewählt. Die anfängliche Abweichung der Marke der Hand liegt bei ca. 10 bis 15 cm. Die Verfolgung des Oberkörpers der Person und der rechten Hand erfolgt im Allgemeinen recht gut. Abbildung 4(b) zeigt die Entwicklung des Fehlers über die Zeit. Der mittlere Fehler über die gesamte Sequenz liegt bei ca. 21 cm, also nicht sehr viel über der initialen Abweichung. Bei einer Entfernung der Person von ca. 3,5 m ist diese Genauigkeit ausreichend für die Erkennung der ausgeführten Handlung mit weiteren Verfahren zur Gestenerkennung, wie bereits in SCHMIDT (2008) gezeigt wurde.

Deutlich ist zu erkennen, dass während der ersten 80 Bilder und zwischen Bild 160 und 250 die Genauigkeit deutlich nachlässt, sich gegen Ende der Sequenz aber wieder verbessert. Tatsächlich hat zwischen Bild 160 und 250 das Körpermodell die Bewegung der Person nicht mehr korrekt verfolgt, da diese am Anfang der Handlung eine schnelle Körperdrehung durchgeführt hat, das Bewegungsmodell dieser aber nicht folgen konnte. Die Position

der Hand wurde trotzdem weiter verfolgt, so dass am Ende der Handlung das Körpermodell die Postur wieder korrekt verfolgen konnte.

Bei genauerer Betrachtung wird deutlich, dass der Fehler bei der Posturschätzung nicht in allen Raumrichtungen gleich verteilt ist, siehe auch Abb. 4(c). Der Fehler in x- und y-Richtung, also parallel zur Bildebene, bleibt während der gesamten Sequenz um 5 bis 10 cm, mit einigen kurzen Ausreißern bis zu 20 cm. Der Fehler in der z-Achse, also in der Tiefe, macht jedoch den größten Anteil am Gesamtfehler aus. Er liegt im Mittel bei 16 cm und bleibt auch über längere Zeiträume über 20 bis zu 40 cm.



**Abb. 4:** Positionsfehler der rechten Hand. (a) Vergleich der Trajektorien: Ground Truth aus photogrammetrischen Marken (grün) und Position der Hand des Modells (blau), (b) absoluter Fehler, (c) Abweichungen einzeln für jede Raumachse.

Problematisch ist die Erkennung von kleinen Bewegungen der Hand, wie zum Beispiel während des Festziehens einer Schraube. Das verwendete Körpermodell kann diesen Bewegungen nur bedingt folgen, da die Hand als starre Fortsetzung des Unterarmes modelliert wurde. Dies ist auch an den wellenförmigen Ausschlägen der Fehlerfunktion erkennbar, die dadurch entstehen, dass das Modell näherungsweise still steht während die Person die Hand in Wirklichkeit hin und her bewegt.

## 5 Zusammenfassung und Ausblick

Auch wenn das monokulare Verfahren im Allgemeinen in der Lage ist, die Postur einer Person in der gegebenen Videosequenz zu verfolgen, so schwankt die Erkennungsgenauigkeit je nach beobachteter Handlung. Generell ist die Frage, wie viel Vorwissen man in das

System integrieren kann. Es ist leicht ersichtlich, dass die Genauigkeit des Modells einen großen Einfluss auf die Ergebnisse der Posturschätzung hat. Hier wurde lediglich ein generisches Körpermodell verwendet, da die genauen Proportionen der Person nicht bekannt waren. Eine Vermessung der Person im Vorfeld oder eine gleichzeitige Schätzung von Postur und Körpermaßen könnten zu genaueren Ergebnissen beitragen. Abschließend bleibt zu erwähnen, dass bei dem Design und der Parametrisierung des Systems auch die Aufgabenstellung berücksichtigt werden sollte. Ist z.B. die Aufgabe die Position der Person zu erkennen, so können andere Merkmale, genutzt werden, als wenn das Ziel die Erkennung der durchgeführten Handlung aufgrund der Trajektorie der Hand ist. In diesem Fall ist die Bewegung der ausführenden Hand relevant, der andere Arm muss aber möglicherweise gar nicht berücksichtigt werden. Auch Wissen über die durchgeführte Handlung kann hilfreich sein, um den Parameterraum gezielter abzusuchen.

Multiokulare Systeme haben bei der hier gestellte Aufgabe einen prinzipiellen Vorteil, da sie Entfernungen direkt messen können oder diese indirekt über einen mehrfachen Modellvergleich für alle Blickwinkel bestimmen können. Steht jedoch nur eine einzelne Kamera zur Verfügung, z.B. weil nur beschränkter Raum zum Einbau vorhanden ist oder weil keine spezialisierten und somit kostspieligen und wartungsintensiven Tiefensensoren verwendet werden sollen, so kann das präsentierte System auch mit nur einer Kamera Tiefeninformationen indirekt durch Ausnutzung des Körpermodells bestimmen. Die erzielte Genauigkeit der Posturerkennung reicht aus, um Rückschlüsse auf die Handlung der Person ziehen zu können.

## 6 Literatur

- Comaniciu, D. & Meer, P. (2002): *Mean shift: a robust approach toward feature space analysis*, IEEE Pattern Analysis and Machine Intelligence, Vol. 24, Nr. 5, S. 603619
- Lu, Z. & Carreira-Perpinan, M. & Sminchisescu, C. (2008): *People tracking with the laplacian eigenmaps latent variable model*, Advances in Neural Information Processing Systems 20, MIT Press, Cambridge, S. 17051712
- Schmidt, J. & Hofemann, N. & Haasch, A. & Fritsch, J. & Sagerer, G. (2008): *Interacting with a mobile robot: Evaluating gestural object references*, Intl. Conference on Intelligent Robots and Systems (IROS), Nice, France
- Schmidt, J. & Kwolek, B. & Fritsch, J. (2006): *Kernel Particle Filter for Real-Time 3D Body Tracking in Monocular Color Images*, IEEE Automatic Face and Gesture Recognition, Southampton, UK, S. 567572
- Sidenbladh, H. (2001): *Probabilistic Tracking and Reconstruction of 3D Human Motion in Monocular Video Sequences*, PhD thesis, KTH Sweden
- Sminchisescu, C. & Triggs, B. (2001): *Covariance scaled sampling for monocular 3d body tracking*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), S. 447454
- Urtasun, R. & Fleet, D. & Fua, P. (2005): *Monocular 3d tracking of the golf swing*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego