Stereo-basierte vs. Monokulare 6-DoF Lagebestimmung unter Verwendung von Punktmerkmalen

Pedram Azad, Tamim Asfour, Rüdiger Dillmann

Institut für Anthropomatik, Karlsruher Institut für Technologie (KIT)
Adenauerring 2, D-76131 Karlsruhe
E-Mail: azad@kit.edu, asfour@kit.edu, dillmann@kit.edu
URL: http://wwwiaim.ira.uka.de

Zusammenfassung

In den letzten Jahren haben Objekterkennungssysteme basierend auf Punktmerkmalen in 2D-Ansichten zunehmend an Bedeutung gewonnen. Derartige Verfahren werden insbesondere für die visuelle Perzeption intelligenter Robotersysteme vielfach eingesetzt. Um Greifaufgaben mit Robotersystemen ausführen zu können, muss zusätzlich zur Erkennung und 2D-Lokalisierung eines Objektes dessen 6-DoF Lage bestehend aus Rotation und Translation im 3D-Raum berechnet werden. Hierzu sind zwei grundlegend verschiedene Ansätze möglich: monokular und stereo-basiert. Während monokulare Ansätze aufgrund ihrer Einfachheit vor allem in Anwendungen der virtuellen Realität populär sind, so haben in der Robotik maximale Genauigkeit und Robustheit höchste Priorität. Im vorliegenden Paper werden die Genauigkeiten des herkömmlichen monokularen und unseres stereo-basierten Ansatzes zur 6-DoF Lagebestimmung basierend auf Punktmerkmalen, sowohl in der Theorie als auch in der Praxis, quantitativ miteinander verglichen. Wir werden zeigen, dass der von uns entwickelte Stereo-basierte Ansatz bezüglich Genauigkeit und insbesondere Robustheit deutlich bessere Ergebnisse erzielt als der monokulare Ansatz.

1 Einleitung

Die exakte Lagebestimmung von Objekten im Raum ist ein wichtige Aufgabe im Bereich des Maschinensehens, insbesondere für Anwendungen der Manipulation mit Robotersystemen. Für die erfolgreiche Ausführung eines Greifvorgangs ist insbesondere auch die exakte Bestimmung der Tiefe grundlegend.

In der jüngeren Vergangenheit hat die Erkennung und 2D-Lokalisierung basierend auf Punktmerkmalen zunehmend an Bedeutung gewonnen und ist

derzeit in der Robotik einer der meist verwendetsten Ansätze. Die bekanntesten Merkmale dieser Art sind die SIFT-Merkmale (Scale Invariant Feature Transform) [9], gefolgt von den neueren SURF-Merkmalen (Speeded Up Robust Features) [5] und auf Regionen basierenden Merkmalen wie die MSER (Maximally Stable Extremal Regions) [12]. Objekterkennungsverfahren, welche solche Merkmale verwenden, operieren üblicherweise auf einer Menge von Merkmalskorrespondenzen. Die Erkennung erfolgt entweder durch einfaches Zählen der Korrespondenzen oder durch Ausnutzung der örtlichen Relationen der Merkmale, welche auch für die Berechnung der 2D-Lokalisierung erforderlich sind. Abb. 1 zeigt ein Beispiel für Korrespondenzen, unter Verwendung der in [2] entwickelten Merkmale.

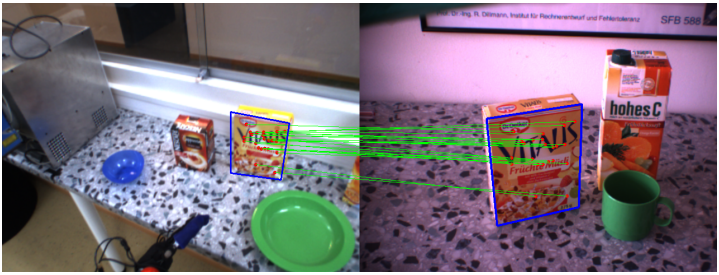

Abbildung 1: Korrespondenzen zwischen aktueller Sicht (links) und eingelernter Sicht (rechts). Aufgrund der Skalierung und Unschärfe des Objektes in der aktuellen Sicht ist das Lokalisierungsergebnis nicht exakt.

Auf der Basis des Ergebnisses der 2D-Lokalisierung kann anschließend die 6-DoF Lage berechnet werden. Herkömmliche, monokulare Ansätze operieren auf 2D-3D Punktkorrespondenzen zwischen 2D-Merkmalspositionen und 3D-Modellpunkten. Die traditionelle Methode hierzu ist der POSIT Algorithmus [7], welcher beispielsweise in [8] für Anwendungen der virtuellen Realität verwendet wird. Ein neuerer Algorithmus, welcher auch für koplanare Punktwolken andwendbar ist, wird in [10] präsentiert. Dieser Algorithmus wurde für die Vergleichsmessungen in den durchgeführten Experimenten verwendet.

All diese Ansätze haben gemeinsam, dass die Lage im Raum auf der Basis eines monokularen Bildes berechnet wird. Dies bedeutet, dass insbesondere die Distanz des Objektes zur Kamera von der Skalierung, d.h. der Größe des Objektes im Bild, abgeleitet wird. Des Weiteren ist die Berechnung der Lage bei Verkippungen des Objektes anfällig gegenüber kleinen Fehlern der 2D-Merkmalspositionen.

Eine Möglichkeit die Genauigkeit der Lagebestimmung zu erhöhen, ist die Anwendung eines kantenbasierten Optimierungsschrittes, welcher die projizierte Kontur des Objektes ausnutzt. Eine solche Optimierung verwendet im

Wesentlichen die gleichen Methoden, die für das kanten- und modellbasierte Tracking von Objekten Einsatz finden (bspw. [11]). Ein hybrider Ansatz, welcher Textur, Kanten- und Farbinformation innerhalb eines *Iterated Extended Kalman Filters* (IEKF) fuisioniert, wird in [13] vorgestellt. Es gilt jedoch zu beachten, dass eine kantenbasierte Verbesserung einer berechneten Lage stets voraussetzt, dass die Kontur des Objektes sich deutlich im Bild abbildet, was oftmals nicht der Fall ist.

Um die o.g. Probleme zu beheben, haben wir einen Ansatz entwickelt, welcher die Vorteile eines kalibrierten Stereokamerasystems ausnutzt. Das Verfahren operiert auf dem Ergebnis der 2D-Erkennung und -Lokalisierung auf der Basis von Merkmalskorrespondenzen, wie erstmalig in [3] beschrieben. Im vorliegenden Paper werden wir sowohl in Theorie als auch in der Praxis zeigen, dass unser Stereo-basierter Ansatz robuster und genauer ist im Vergleich zu herkömmlichen monokularen Ansätzen basierend auf 2D-3D Punktkorrespondenzen. Es gilt zu beachten, dass es weder eine akzeptierte Tatsache noch offensichtlich ist, dass monokulare Lageschätzung auf der Basis von 2D-3D Punktkorrespondenzen zu einem schlechteren Ergebnis führt als Stereo-basierte Lageschätzung[1].

In [6] wird ein verwandter Ansatz für das 3D-Tracking eines Objektes vorgestellt, welcher den KLT-Algorithmus [14] für das Verfolgen von Merkmalen verwendet um die Rechenzeit zu reduzieren. Es werden ein monokularer Ansatz unter Verwendung von 2D-3D Punktkorrespondenzen und ein Stereo-basierter Ansatz zur Lagebestimmung vorgestellt. Obwohl experimentell gezeigt wird, dass der Stereo-basierte Ansatz genauere Ergebnisse liefert, wird keine detaillierte Analyse durchgeführt.

Unsere Messungen zeigen, in welchen Situationen der monokulare und der Stereo-basierte Ansatz vergleichbare Ergebnisse liefern, und in welchen Fällen sich die Genauigkeit des monokularen Ansatzes verschlechtert. Insbesondere werden wir zeigen, dass der monokulare Ansatz für planare Objekte bei Verkippungen Instabilitäten aufweist und dass der Stereo-basierte Ansatz für weite Entfernungen des Objekts eine wesentlich höhere Tiefengenauigkeit erzielt. Darüberhinaus erzielt unser Stereo-basierter Ansatz maximale Genauigkeit durch das Einpassen eines 3D-Modells anstatt die Lage ausschließlich aus Punktkorrespondenzen abzuleiten.

In Abschnitt 2 werden die maximal erzielbare Genauigkeiten des monokularen und Stereo-basierten Ansatzes theoretisch miteinander verglichen. Unser Stereo-basierter Ansatz zur 6-DoF Lagebestimmung wird in Abschnitt 3 beschrieben. Der monokulare und Stereo-basierte Ansatz werden in Simulation

[1] Nicht jeder Stereo-basierte Ansatz erzielt zwangsläufig eine höhere Genauigkeit. Beispielsweise ist der Ansatz die 2D-Lokalisierung im linken und rechten Kamerabild getrennt durchzuführen und die Ergebnisse in ein 3D-Ergebnis zu fusionieren - sowohl in theoretischer als auch in praktischer Hinsicht – suboptimal in puncto Genauigkeit.

und in realen Experimenten in Abschnitt 4 miteinander verglichen. Abschließend wird eine Zusammenfassung in Abschnitt 5 gegeben.

2 Genauigkeitsbetrachtungen

In diesem Abschnitt wird die maximal erzielbare Genauigkeit der Lagebestimmung basierend auf 2D-3D Korrespondenzen mit auf Stereo-Triangulation basierenden Verfahren verglichen. Für die Kameraparameter wurden die Werte der realen vorherrschenden Verhältnisse des weitwinkligen Kamerapaares des humanoiden Roboters ARMAR-III [1] herangezogen. Als Aufgabe wurde die Lokalisierung eines Objekts in Manipulationsentfernung von ca. 75 cm mit dem Ziel des anschließenden Greifens definiert. Die weitwinkligen Linsen mit einer Brennweite von 4 mm ergeben gemäß dem Ergebnis des Kalibriervorgangs ca. $f = f_x = f_y = 530$ (Pixel). Das Stereokamerasystem besitzt einen Kameraabstand von $b = 90$ mm; die Hauptachsen der beiden Kameras werden als parallel angenommen.

Wie in [2] gezeigt wird, führt ein Pixel-Fehler von Δ Pixel zu einem relativen Fehler der zu bestimmenden z_c-Koordinate von:

$$\frac{z_c(u)}{z_c(u+\Delta)} - 1 = \frac{\Delta}{u}\,. \tag{1}$$

Dies zeigt, dass der Fehler – zusätzlich zum Pixel-Fehler – von der projizierten Größe des Objektes abhängt: Je größer die projizierte Größe u, umso kleiner der Fehler. Für die Berechnung der Lage auf der Basis von Merkmalspunkten steht u im Optimalfall in Bezug zu der weitesten Entfernung zweier Merkmalspunkte. Für ein Objekt, dessen Merkmalspaar mit der weitesten Entfernung eine Distanz von 100 mm besitzt, ist $u = \frac{f \cdot x_c}{z_c} \approx 70$, unter der Annahme dass die Objektoberfläche und die Bildebene parallel verlaufen. Ein Pixelfehler von $\Delta = 1$ würde bereits zu einem absoluten Fehler der z_c-Koordinate von $75\,\text{cm} \cdot \frac{1}{70} \approx 1\,\text{cm}$ führen, unter ansonsten optimalen Bedingungen.

In einem realistischen Szenario jedoch liegen die Objekte meist mit Verkippungen vor, welche ein verzerrtes Abbild des Objektes zur Folge haben. Diese Verzerrungen verursachen nicht nur ein kleinere projizierte Fläche des Objektes, sondern auch einen größeren Fehler der Orte der Merkmalspunkte. Eine projizierte Größe von 50 Pixel und ein effektiver Pixelfehler von $\Delta = 1.5$ Pixel würde bereits zu einem absoluten Fehler von über 2 cm im betrachteten Beispiel führen. Es gilt zu beachten, dass die Tiefengenauigkeit nicht nur von den Pixelfehlern im aktuell vorliegenden Bild, sondern auch vom Trainingsbild abhängt, da die Tiefe relativ zum Trainingsbild bestimmt wird.

Im Gegensatz dazu, wird bei Ausnutzung eines kalibrierten Stereokamerasystems die Tiefe ausschließlich auf der Basis der aktuellen Ansicht berechnet. Wie in [2] gezeigt wird, führt ein Fehler von Δ Pixel in der Disparität zu einem

relativen Fehler der zu bestimmenden z_c-Koordinate von:

$$\frac{z_c(d)}{z_c(d+\Delta)} - 1 = \frac{\Delta}{d}\,, \tag{2}$$

wobei d die Disparität zwischen dem linken und dem rechten Kamerabild bezeichnet. Gleichung (2) zeigt, dass der Fehler nicht von der projizierten Größe des Objektes abhängt, sondern stattdessen von der Disparität d: Je größer die Disparität, umso kleiner der Fehler. Für das spezifizierte Stereokamerasystem beträgt die Disparität $d = \frac{f \cdot b}{z_c} \approx 64$. Die Korrespondenzen zwischen dem linken und rechten Kamerabild können für die Merkmalspunkte mit Subpixel-Genauigkeit berechnet werden. In der Praxis kann eine Subpixel-Genauigkeit von mindestens 0.5 Pixel angenommen werden. Gemäß Gleichung (2) ergibt dies einen Gesamtfehler von nur $75\,\mathrm{cm} \cdot \frac{0.5}{64} \approx 0.6\,\mathrm{cm}$.

Die vorgestellten theoretischen Berechnungen zeigen, dass durch Ausnutzung des Stereo-Sehens die Genauigkeit um Faktor 2–3 gesteigert werden kann. Die tatsächlichen Fehler jedoch, welche bei der Lageschätzung auf der Basis von 2D-3D Punktkorrespondenzen entstehen, können nicht durch theoretische Formeln ausgedrückt werden. Die Genauigkeit und Stabilität solcher Ansätze hängt grundlegend von der räumlichen Verteilung der Merkmalspunkte und deren Genauigkeit ab.

3 6-DoF Lageschätzung

Herkömmliche Ansätze zur Lagebestimmung, welche auf 2D-3D Punktkorrespondenzen basieren, können keine ausreichende Genauigkeit und Robustheit erreichen. Sie tendieren insbesondere dazu instabil zu werden, wenn die effektive Auflösung des Objektes niedrig ist und somit auch die relative Genauigkeit der 2D Positionen der Merkmalspunkte. Nachfolgend stellen wir unseren Ansatz vor, welcher die Vorteile eines kalibrierten Stereokamerasystems ausnutzt. Wir werden zeigen, dass unser Ansatz eine signifikant höhere Robustheit und Genauigkeit aufweist und auch bei geringer Auflösung des Objekts erfolgreich ist.

Die Idee ist, zunächst eine dünn besetzte 3D-Punktwolke für die 2D-Fläche zu berechnen, welche durch die Transformation der Kontur aus dem Trainingsbild in die aktuelle Sicht gegeben ist. Diese Transformation liegt als Ergebnis der 2D-Lokalisierung vor [4, 2]. Bei einem gegebenen 3D-Modell, kann dieses mit der berechneten Punktwolke registriert und auf diese Weise eine 6-DoF Lage berechnet werden. Der allgemeine Ansatz ist in Algorithmus 1 zusammengefasst.

In Schritt 4 aus Algorithmus 1 sind zwei Varianten möglich: Das Einpassen einer analytisch formulierten 3D-Repräsentation (oder ein hoch aufgelöstes Polygon-Modell) in die Punktwolke oder die Berechnung der Lage auf Basis

Algorithm 1 BerechneLage(I_l, I_r, C) $\rightarrow R, \boldsymbol{t}$

1. Bestimme die Menge an Eckpunkten (engl. *interest points*) innerhalb der berechneten 2D-Kontur C des Objekts im linken Kamerabild I_l.
2. Für jeden berechneten Punkt, bestimme die Korrespondenz im rechten Kamerabild I_r durch Berechnung der *Zero Normalized Cross Correlation* (ZNCC) entlang der Epipolarlinie.
3. Berechne den 3D-Punkt für jede Korrespondenz durch Stereo-Triangulation.
4. Passe ein 3D-Modell des Objekts in die berechnete Puntkwolke durch Anwendung eines Registrierungsverfahrens ein und gebe die auf diese Weise berechnete Rotation R und Translation $\boldsymbol{t}$ als Ergebnis zurück.

von expliziten 3D-3D Punktkorrespondenzen. Um die erste Variante einsetzen zu können, muss ein geometrisches 3D-Modell des Objektes vorliegen. Für die zweite Variante müssen 3D-Punkte für die Merkmalspunkte aus der Trainingsansicht auf die selbe Art und Weise wie bei der Erkennung berechnet werden, d.h. durch Bestimmung von 2D-2D Punktkorrespondenzen und Stereo-Triangulation. Eine Menge von 3D-3D Punktkorrespondenzen ist dann automatisch durch die gefilterte Menge der 2D-2D Punktkorrespondenzen gegeben, welche als Ergebnis der 2D-Lokalisierung vorliegt.

Falls anwendbar, sollte die erste Variante bevorzugt werden, da sie nicht von der Genauigkeit bzw. Reproduzierbarkeit der Positionen der Merkmalspunkte zwischen aktueller Sicht und Trainingsansicht abhängt. Jedoch auch die zweite Variante ist genauer als der herkömmliche monokulare Ansatz, da sie nicht die Instabilitäten aufweist, welche typisch für die Lagebestimmung auf der Basis von 2D-3D Punktkorrespondenzen sind.

Für den Fall von quaderförmigen Objekten, wie sie für die durchgeführeten Experimente eingesetzt wurden, kann die Kontur durch die vier Eckpunkte der Vorderfläche berechnet werden. Es wird Variante 1 verwendet und für die Registrierung eine 3D-Ebene als Repräsentation für die Vorderfläche eingepasst. Details hierzu sind in [2] beschrieben.

4 Experimentelle Evaluation

Nachfolgend werden die Genauigkeiten des monokularen und des vorgestellten Stereo-basierten Ansatzes in mehreren Experimenten miteinander verglichen. Für die Erkennung und 2D-Lokalisierung werden die Merkmale und das Verfahren aus [4] verwendet. Das System wurde mithilfe des Integrating

Vision Toolkit[2] (IVT) implementiert. Die Firma Keyetech[3] bietet hochoptimierte Implementierungen von Funktionen des IVT (z.B. Harris Eckendetektor innerhalb von 4,2 ms für Bilder der Größe 640×480).

In den ersten Experimenten wurde das in Abschnitt 2 spezifizierte Stereokamerasystem simuliert, sodass die Fehler unter optimalen Bedingungen in Bezug zu *ground truth* Information berechnet werden konnten. In Abb. 2 sind die Fehler der z-Koordinate dargestellt, da diese die Schwachstelle des monokularen Ansatzes aufzeigen. Hierzu wurde für jedes Diagramm das Objekt entlang bzw. um jeweils einen einzelnen Freiheitsgrad bewegt bzw. rotiert. Zusätzlich wurden 1.000 zufällige Objektlagen evaluiert; die Ergebnisse sind in Abb. 3 dargestellt.

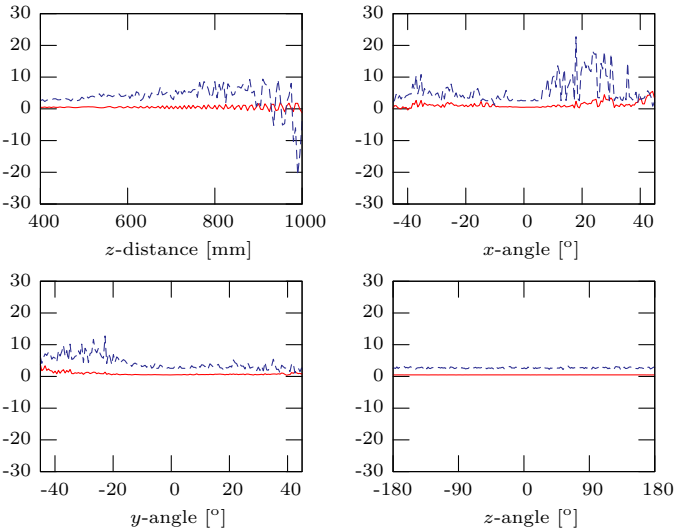

Abbildung 2: Ergebnisse des z-Fehlers in Abhängigkeit jeweils eines einzelnen Freiheitsgrads. Die durchgezogene Linie zeigt das Ergebnis des vorgestellten Stereo-basierten Ansatzes, die unterbrochene Linie die des monokularen Ansatzes.

Abb. 4 zeigt eine Situation, in der der monokulare Ansatz instabil wird. In Tabelle 1 sind die Standardabweichungen für eine Bildsequenz bestehend aus 100 Aufnahmen für ein reales Experiment mit einem statischen Objekt aufgeführt. Wie zu sehen ist, beträgt die Standardabweichung der z-Koordinate 1,52 mm unter Verwendung des monokularen Ansatzes im Vergleich zu 0,39 mm

[2] `http://ivt.sourceforge.net`
[3] `http://www.keyetech.de`

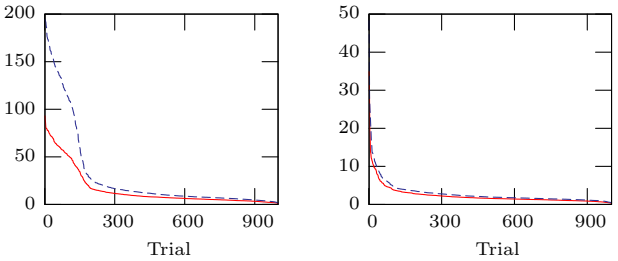

Abbildung 3: Genauigkeit der 6-DoF Lageschätzung für 1.000 zufällige Versuche; die Fehler sind absteigend sortiert. Die durchgezogene Linie zeigt den durchschnittlichen Fehler, die unterbrochene Linie den maximalen Fehler. Der 3D-Fehler wurde auf der Basis von hochauflösend abgetasteten Oberflächenpunkten berechnet. Links: unter Verwendung des monokularen Ansatzes. Rechts: unter Verwendung der Stereo-basierten Ansatzes. Der Leser sei auf die unterschiedliche Skalierung der vertikalen Achsen hingewiesen.

unter Verwendung des vorgestellten Stereo-basierten Ansatzes. Für weitere Genauigkeitsmessungen sei auf [2] verwiesen.

Abbildung 4: Ergebnis der Lagebestimmung für eine Beispiel-Szene. Links: unter Verwendung des monokularen Ansatzes. Rechts: unter Verwendung des Stereo-basierten Ansatzes.

Die Laufzeit der Lagebestimmung beträgt ca. 6 ms für ein einzelnes Objekt für den in Abschnitt 2 spezifizierten Aufbau, unter Verwendung eines 3 GHz Intel Pentium 4. Der einzige rechenaufwändige Schritt ist die Bestimmung der Korrespondenzen durch Korrelation. Die Laufzeit kann auf einfache Art und Weise durch Einbezug der Korrelationsergebnisse benachbarter Merkmalspunkte deutlich reduziert werden.

	x	y	z	θ_x	θ_y	θ_z
Proposed method	0,23	0,42	0,39	0,066	0,17	0,10
Conventional method	0,24	0,038	1,52	0,17	0,29	0,13

Tabelle 1: Standardabweichungen der geschätzen Lagen für ein statisches Objekt über eine Bildsequenz bestehend aus 100 Aufnahmen. Die Einheiten sind [mm] bzw. [°]. Es wurde eine Situation ausgewählt, in welcher der monokulare Ansatz *nicht* instabil wird.

5 Disskussion und Ausblick

Wir haben monokulare 6-DoF Lageschätzung auf der Basis von 2D-3D Punktkorrespondenzen mit unserem Stereo-basierten Ansatz verglichen. Nach einer Disskussion beider Ansätze wurde gezeigt, dass der Stereo-basierte Ansatz eine signifikant höhere Robustheit und Genauigkeit aufweist. Die größten Abweichungen zwischen den beiden Ansätzen konnten in der z-Koordinate beobachtet werden.

In der näheren Zukunft planen wir den vorgestellten Stereo-basierten Ansatz für beliebig geformte Objekten zu anzuwenden. Insbesondere soll evaluiert werden, welche Verbesserung sich durch das Einpassen eines 3D-Modells zusätzlich zur Lagebestimmung rein basierend auf expliziten 3D-3D Punktkorrespondenzen erzielen lässt.

Danksagung

Die im diesem Paper beschriebene Arbeit wurde im Rahmen der durch die Europäische Kommission finanzierten EU-Projekte PACO-PLUS (IST-FP6-IP-027657) und GRASP (IST-FP7-IP-215821) sowie im Rahmen des durch die Deutsche Forschungsgemeinschaft (DFG) finanzierten Sonderforschungsbereichs “Humanoide Roboter” (SFB 588) durchgeführt.

Literatur

[1] T. Asfour, K. Regenstein, P. Azad, J. Schröder, N. Vahrenkamp, and R. Dillmann. ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 169–175, Genova, Italy, 2006.

[2] P. Azad. *Visual Perception for Manipulation and Imitation in Humanoid Robots*. PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, 2008.

[3] P. Azad, T. Asfour, and R. Dillmann. Stereo-based 6D Object Localization for Grasping with Humanoid Robot Systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 919–924, San Diego, USA, 2007.

[4] P. Azad, T. Asfour, and R. Dillmann. Combining Harris Interest Points and the SIFT Descriptor for Fast Scale-Invariant Object Recognition. In *submitted to IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, St. Louis, USA, 2009.

[5] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *European Conference on Computer Vision (ECCV)*, pages 404–417, Graz, Austria, 2006.

[6] C. Choi, S.-M. Baek, and S. Lee. Real-time 3D Object Pose Estimation and Tracking for Natural Landmark Based Visual Servo. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3983–3989, Nice, France, 2008.

[7] D. F. DeMenthon and L. S. Davis. Model-Based Object Pose in 25 Lines of Code. In *European Conference on Computer Vision (ECCV)*, pages 123–141, Santa Margherita Ligure, Italy, 1992.

[8] V. Lepetit, L. Vacchetti, D. Thalmann, and P. Fua. Fully Automated and Stable Registration for Augmented Reality Applications. In *International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 93–102, Tokyo, Japan, 2003.

[9] D. G. Lowe. Object Recognition from Local Scale-Invariant Features. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1517, Kerkyra, Greece, 1999.

[10] C.-P. Lu, G. D. Hager, and E. Mjolsness. Fast and Globally Convergent Pose Estimation from Video Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(6):610–622, 2000.

[11] E. Marchand, P. Bouthemy, F. Chaumette, and V. Moreau. Robust Real-Time Visual Tracking using a 2D-3D Model-based Approach. In *IEEE International Conference on Computer Vision (ICCV)*, pages 262–268, Kerkyra, Greece, 1999.

[12] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *British Machine Vision Conference (BMVC)*, volume 1, pages 384–393, London, UK, 2002.

[13] G. Taylor and L. Kleeman. Fusion of Multimodal Visual Cues for Model-Based Object Tracking. In *Australasian Conference on Robotics and Automation (ACRA)*, Brisbane, Australia, 2003.

[14] C. Tomasi and T. Kanade. Detection and Tracking of Point Features. Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburgh, USA, 1991.