Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Anwendungen 1 WS08/09 - Ausarbeitung

Benjamin Wagner

3D-Objekterkennung im Kontext eines Assistenzroboters

Betreuende Prüfer
Prof. Dr. rer. nat. Stephan Pareigis
Prof. Dr. Birgit Wendholt

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Benjamin Wagner

Thema der Ausarbeitung

3D-Objekterkennung im Kontext eines Assistenzroboters

Stichworte

SIFT, Kameraparameter, Objektpose, Bildtrajektorie

Kurzzusammenfassung

In dieser Ausarbeitung werden aktuelle Verfahren zur Erkennung von 3D-Objekten vorgestellt. Für jedes Verfahren wird die Modellierung von Referenzmodellen, die Erkennung von Objekten in Bildern und die Bestimmung der Objektpose beschrieben. Eine der Vorgehensweisen wird ausgewählt, um zukünftig als Grundlage für das autonome Greifen von Gegenständen durch einen Robotergreifarm verwendet zu werden. Die entscheidenden Kriterien bei der Auswahl eines Verfahrens sind die Schnelligkeit und die Qualität der Wiedererkennung von Objekten in Bildern.

Inhaltsverzeichnis

1 Problemstellung

Ein Robotergreifarm, der über eine Videokamera an seiner Hand verfügt, soll Objekte autonom identifizieren und greifen können. Dazu wird der Katana-Greifarm aus Bild 1.1 eingesetzt, bei dem es sich um einen 6-Gelenk-Knickarm-Roboter handelt. Die Grauwertbilder der Videokamera werden an einen externen PC gesendet und dort ausgewertet. Auf Basis der ausgewerteten Daten soll der PC den Greifarm steuern.

Bild 1.1: Katana-Greifarm

Die vorliegende Problemstellung ist auf nachfolgend aufgeführte Aufgaben beschränkt.

- Bereitstellung von Referenzdaten für zu erkennende Objekte,
- bildabhängige Identifikation von Objekten aus einer beliebigen Perspektive,
- Bestimmung der Pose von identifizierten Objekten.

Die *Objektpose* beinhaltet die Position und die Orientierung eines Objekts ausgehend von einer Kamera. Die ermittelte Objektpose ermöglicht es einen identifizierten Gegenstand zu greifen. In Abschnitt 2.2 wird näher auf den Begriff Objektpose eingegangen. An dieser Stelle sei darauf hingewiesen, dass die Aufgabenstellung nicht den Greifvorgang beinhaltet.

Es besteht die Anforderung, dass ein zu erkennender Gegenstand nicht im Bild vorhanden sein muss. Eine Suche nach einem Objekt, durch die Bewegung des Greifarms, soll zukünftig möglich sein. Das ist jedoch nicht Teil der hier vorliegenden Problemstellung. Weiterhin soll nicht berücksichtigt werden, dass ein Objekt in einem Bild doppelt vorkommen kann oder partiell verdeckt sein kann. Außerdem ist die Erkennung von deformierbaren Gegenständen nicht beabsichtigt.

Eine weitere Anforderung an das System ist, dass nur ein Objekt zur Zeit identifiziert werden können soll. Das bedeutet, dass die aktuellen Bilddaten für die Dauer eines Erkennungsversuchs nur mit einer Objektreferenz verglichen werden müssen. Dies ist von Vorteil für die Anforderung, dass die Dauer des Vergleichs einer aktuellen Szene mit den Referenzdaten so kurz wie möglich sein soll. Darunter soll aber nicht die Qualität der Erkennung von Objekten leiden.

Das Ziel ist es einen Greifarm für Assistenzaufgaben einzusetzen. Aus diesem Grund ist der verwendete Arm im Verhältnis zu einem Mensch klein und nicht stark, damit Personen in der direkten Umgebung nicht verletzt werden können. Der Katana-Greifarm aus Bild 1.1 hat eine Nutzlast von 500 Gramm. Somit können nur kleine leichte Gegenstände gegriffen werden. Im Folgenden werden Beispiele für konkrete Anwendungsfälle aufgelistet.

- autonomes Greifen von Werkstücken in Gefahrenzonen,
- Bar-Roboter, welcher auf Wunsch Getränke serviert,
- Assistenz-Roboter im Pflegebereich für das Zureichen von Gegenständen.

Die hier vorliegende Problematik fällt in den Bereich der 3D-Bildverarbeitung/3D-Objekterkennung. In den folgenden Kapiteln werden aktuelle Verfahren und Ideen zur Lösung des beschriebenen Problems vorgestellt und diskutiert.

In Gliederungspunkt 2 werden essentielle Grundlagen eingeführt, die in den Verfahren in Kapitel 3 verwendet werden. Die Erkenntnisse aus Abschnitt 3 werden in Kapitel 4 zusammengefasst. Anschließend wird entschieden welche Vorgehensweise für die Lösung der Problemstellung unter Berücksichtigung der gegebenen Anforderungen prädestiniert ist. In Gliederungspunkt 5 wird erläutert welche Aufgaben zur Umsetzung des gewählten Verfahrens gelöst werden müssen und welche Risiken dabei auftreten können.

2 Grundlagen

2.1 Scale Invariant Feature Transform (SIFT)

Die bildabhängige Identifikation von Gegenständen passiert in den Verfahren in Abschnitt 3 auf Basis eines Vergleichs von *markanten Bildpunkten* eines aufgenommenen Bildes mit den markanten Bild- oder Raumpunkten eines Referenzmodells. Bei markanten Bildpunkten handelt es sich um visuell deutlich unterscheidbare Pixel. In Gliederungspunkt 2.2 wird näher auf markante Raumpunkte eingegangen.

Zur Beschreibung von markanten Bildpunkten können SIFT-Merkmalsvektoren verwendet werden. Diese Merkmalsvektoren sind dazu geeignet, um einen Vergleich von Objekten in Bildern durchzuführen, die aus unterschiedlichen Perspektiven erstellt wurden (vgl. LOWE (2004)). Zusätzlich dürfen die Szenen in der ein Objekt auftaucht verschieden sein. Die durch SIFT-Merkmalsvektoren beschriebenen Bildpunkte sind sehr unterscheidbar. Wenn ein markanter Bildpunkt X eines Objektes mit vielen markanten Bildpunkten, die zu anderen Objekten gehören, verglichen wird, dann wird mit hoher Wahrscheinlichkeit keiner der Bildpunkte mit Bildpunkt X übereinstimmen. SIFT-Merkmalsvektoren sind invariant hinsichtlich der Skalierung und der Rotation, sowie robust bei der Veränderung des Aufnahmestandpunktes und der Beleuchtung. Auf diese Eigenschaften wird am Ende dieses Abschnitts näher eingegangen.

Es folgt ein Beispiel, um die Anwendung von SIFT-Merkmalsvektoren darzustellen. In Bild 2.1 a und b wird das Hamburger Rathaus aus zwei verschiedenen Aufnahmestandpunkten gezeigt. Die Merkmalsvektoren wurden für beide Bilder berechnet und durch weiße Kreuze in den Bildern markiert. In Bild 2.2 sind die übereinstimmenden markanten Bildpunkte aus Bild 2.1 a und b

durch weiße Linien miteinander verbunden. Man kann erkennen, dass viele der markanten Bildpunkte aus Bild 2.1 a auch in Bild 2.1 b vorkommen. Diese Punkte stimmen, soweit es auf Grund der Menge der weißen Linien ersichtlich ist, auch korrekt überein.

In LANGE (2008) wurde allerdings anhand von anderen Beispielen gezeigt, dass auch falsche Übereinstimmungen in geringem Maße vorkommen können. Aus dieser Erkenntnis kann man schließen, dass für die erfolgreiche Wiedererkennung eines Objektes in einem Bild genügend markante Bildpunkte vorhanden sein müssen. Anhand des hier gezeigten Beispiels lässt sich eine weitere Eigenschaft von SIFT-Merkmalsvektoren erkennen. Bevor die markanten Bildpunkte eines Bildes mit Referenzdaten verglichen werden können, muss das Bild nicht erst segmentiert werden (vgl. MIKOLAJCZYK UND SCHMID (2005)). Laut NISCHWITZ ET AL. (2007) handelt es sich bei einer Segmentierung um die Aufteilung eines Bildes in einzelne Bereiche. Man benötigt also keinen Verarbeitungsschritt, um ein Objekt in einem Bild von seinem Hintergrund zu trennen. Somit kann Rechenzeit bei dem Versuch einer Erkennung eines Objekts eingespart werden.

(a) (b)

Bild 2.1: Das Hamburger Rathaus aus zwei verschiedenen Perspektiven (Quelle LANGE (2008))

Bild 2.2: Übereinstimmende markante Bildpunkte aus Bild 2.1 a und b. (Quelle LANGE (2008))

Laut LOWE (2004) kann der Vergleich eines Bildes mit zwei Referenzbildern auf einem Standard-PC in weniger als 0,3 Sekunden ausgeführt werden. Außerdem wird erwähnt, dass die Invarianz hinsichtlich der Rotation für 3D-Objekte 30 Grad in jede Richtung beträgt. In LANGE (2008) wird das SIFT-Verfahren empirisch analysiert, um die Grenzen der Invarianz hinsichtlich der Skalierung und der Rotation zu ermitteln. Weiterhin wird die Robustheit des Verfahrens bei

Veränderung der Beleuchtung untersucht. Es wurde festgestellt, dass markante Bildpunkte eines Objekts, das bis zur halben Größe im Vergleich zum Referenzbild vorhanden ist, vielversprechend wiedererkannt werden können. Außerdem wurde ermittelt, dass eine geringfügige Veränderung der Beleuchtung einen marginalen Einfluss auf die Wiedererkennung von markanten Bildpunkten hat.

Zur Beschreibung von markanten Bildpunkten gibt es neben dem SIFT-Verfahren noch andere Vorgehensweisen. In MIKOLAJCZYK UND SCHMID (2005) werden verschiedene Verfahren zur Beschreibung von markanten Bildpunkten miteinander verglichen, die sich im Kontext der Objekterkennung bereits bewährt haben. Dabei stand speziell die Qualität der Wiedererkennung im Vordergrund. Es hat sich herausgestellt, dass das SIFT-Verfahren zu den besten Methoden gehört. Nur das GLOH-Verfahren, welches auf SIFT basiert, erzielt geringfügig bessere Ergebnisse. Allerdings wird in BAY ET AL. (2006) aufgeführt, dass GLOH langsamer ist als SIFT. Die vorgestellten Verfahren zur 3D-Objekterkennung in Gliederungspunkt 3 verwenden das SIFT-Verfahren oder Methoden die auf dem gleichen Prinzip basieren.

Eine Schwäche der SIFT-Merkmalsvektoren ist, wie auch in LEE ET AL. (2006) erwähnt, dass Objekte ohne Texturen nicht wiedererkannt werden können. Diese Eigenschaft ist in Bild 2.1 a und b zu erkennen. Der Himmel besitzt nur wenig Texturen und somit wird für diesen Bereich nur eine geringe Anzahl von markanten Bildpunkten detektiert.

2.2 Kamerakalibrierung

2.2.1 Einleitung

Die Abbildung einer dreidimensionalen Szene auf ein Bild mittels einer Kamera wird als Projektion bezeichnet, welche auf den intrinsischen und extrinsischen Parametern der verwendeten Kamera basiert (vgl. HARTLEY UND ZISSERMAN (2003)). Die Ermittlung dieser Parameter nennt man *Kamerakalibrierung*. Bei den *äußeren Kameraparametern* handelt es sich um die Position und um die Orientierung einer Kamera in einem Weltkoordinatensystem. Die äußeren Kameraparameter können zusammengefasst auch als *Kamerapose* bezeichnet werden. Ein Weltkoordinatensystem enthält Raumpunkte $\in \Re^3$ die bei einer Projektion mittels einer Kamera auf Bildpunkte $\in \Re^2$ abgebildet werden. Eine solche Projektion kann durch ein mathematisches Modell der verwendeten Kamera beschrieben werden.

Im Folgenden wird ein Modell einer einfachen Lochkamera vorgestellt. Anschließend wird grob erläutert wie man die extrinsischen Kameraparameter auf Basis des Modells bestimmen kann.

2.2.2 Modell einer Lochkamera

Das Modell einer grundlegenden Lochkamera ist laut HARTLEY UND ZISSERMAN (2003) das einfachste Kameramodell, welches allerdings teilweise die Projektionseigenschaften von modernen Digitalkameras beschreibt.

Es wird eine Zentralprojektion von Raumpunkten $X = (X_C, Y_C, Z_C)^T$ eines Kamerakoordinatensystems auf Bildpunkte $x = (x_c, y_c)^T$ eines Bildkoordinatensystems betrachtet. Das Projektionszentrum ist der Ursprung des Kamerakoordinatensystems und wird auch als Kame-

razentrum oder optisches Zentrum bezeichnet. Lichtstrahlen die durch das Projektionszentrum in die Kamera gelangen sorgen für eine Abbildung von Raumpunkten auf Bildpunkte. Dieser Vorgang wird in Grafik 2.3 verdeutlicht. Dabei ist allerdings zu beachten, dass sich die Bildfläche zu Darstellungszwecken vor dem optischen Zentrum befindet. In Wirklichkeit befindet sich die Bildebene mit dem Abstand der Größe f hinter dem Kamerazentrum und eine aufgenommene Szene wird in einem Bild auf dem Kopf dargestellt. f ist die Brennweite und stellt einen *inneren Kameraparameter* dar.

Mit Hilfe des Strahlensatzes ergibt sich Gleichung 1 für die Zentralprojektion in euklidischen Koordinaten. Gleichung 1 lässt sich kompakter in homogenen Koordinaten in Formel 2 darstellen. Der Zusammenhang zwischen euklidischen und homogenen Koordinaten wird in HARTLEY UND ZISSERMAN (2003) vertieft erläutert.

$$x = (x_c, y_c)^T = (f * \frac{X_C}{Z_C}, f * \frac{Y_C}{Z_C})^T \tag{1}$$

$$\begin{pmatrix} x_c \\ y_c \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} * \begin{pmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{pmatrix} \tag{2}$$

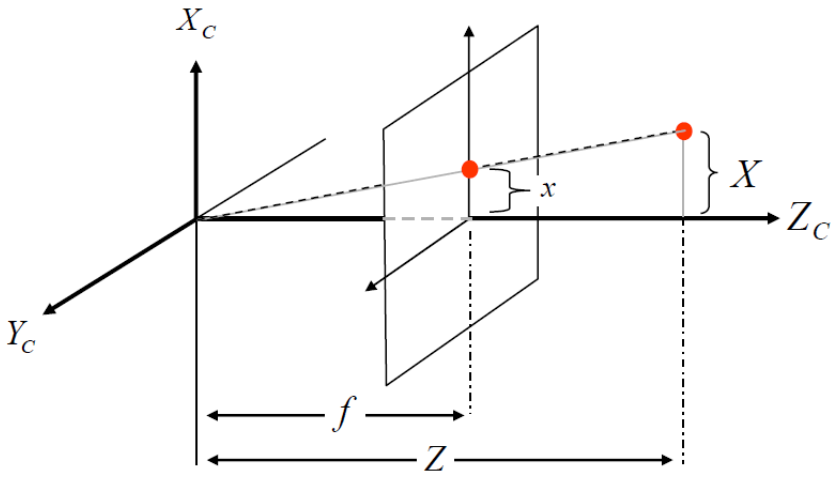

Bild 2.3: Zentralprojektion von Raumpunkten $X = (X_C, Y_C, Z_C)^T$ auf Bildpunkte $x = (x_c, y_c)^T$ (Quelle MEISEL (2008))

Der Punkt in dem sich die optische Achse Z_C in Grafik 2.3 mit der Bildebene schneidet wird Bildhauptpunkt genannt. Dies ist der Ursprung des Bildkoordinatensystems. In MEISEL (2008) wird erläutert, dass die Koordinaten eines Bildpunktes vorteilhaft in den Koordinaten des Bildverarbeitungssystems angegeben werden können. Somit gibt es nur positive Bildpunktindizes. Der Ursprung des Bildkoordinatensystems liegt dann in der linken oberen Ecke und um die Pixel p_x, p_y versetzt zum Bildhauptpunkt. Die Bildpunkte $x = (x_c, y_c)^T$ werden jetzt als Bestandteil des Bildkoordinatensystems des bildverarbeitenden Systems mit $x = (x, y)^T$ bezeichnet. Somit ergibt sich aus Gleichung 2 die erweiterte Formel 3. p_x und p_y gehören neben f zu den *internen Kameraparametern*.

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & p_x & 0 \\ 0 & -f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} * \begin{pmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{pmatrix} \tag{3}$$

Die projizierten Raumpunktkoordinaten sind in der Realität nicht Teil des Kamerakoordinatensystems, sondern des Weltkoordinatensystems (vgl. HARTLEY UND ZISSERMAN (2003)). Es muss also die Verschiebung und die Verdrehung des Kamerakoordinatensystems ausgehend vom Ursprung des Weltkoordinatensystems berücksichtigt werden. Aus diesem Grund muss Gleichung 3 um den Translationsvektor $t = (t_x, t_y, t_z)^T$ erweitert werden, der die Verschiebung des Kamerazentrums vom Ursprung des Weltkoordinatensystems in Weltkoordinaten angibt. Weiterhin muss in Formel 3 die 3 x 3 Rotationsmatrix R berücksichtigt werden. Die 9 Koeffizienten der Rotationsmatrix sind funktional abhängig von den 3 Rotationswinkeln α, β und γ (vgl. MEISEL (2008)). Die Raumpunkte $X = (X_C, Y_C, Z_C)^T$ werden nun als Teil des Weltkoordinatensystems mit $X = (X, Y, Z)^T$ bezeichnet. Aus Gleichung 3 folgt somit Formel 4. Gleichung 4 kann in Kurzform auch als $x = P * X$ dargestellt werden, wobei P als Projektionsmatrix bezeichnet wird.

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} f & 0 & p_x & 0 \\ 0 & -f & p_y & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} * \begin{pmatrix} \mathbf{R} & \mathbf{-R * t} \\ \mathbf{0} & 1 \end{pmatrix} * \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{4}$$

2.2.3 Kalibrierung der extrinsischen Kameraparameter

Die Rotationsmatrix R und der Translationsvektor $t = (t_x, t_y, t_z)^T$ aus Abschnitt 2.2.2 beinhalten folgende extrinsische Kameraparameter, die zusammengefasst auch als Kamerapose bezeichnet werden.

- Position $t = (t_x, t_y, t_z)^T$ einer Kamera in einem Weltkoordinatensystem,
- Orientierung einer Kamera in einem Weltkoordinatensystem, die durch die Rotationswinkel α, β und γ festgelegt ist.

Um ein identifiziertes Objekt greifen zu können, muss die Kamerapose bekannt sein. Die Voraussetzung für die Ermittlung der Kamerapose ist, dass für einen zu identifizierenden Gegenstand zum Beispiel ein 3D-Objektmodell vorhanden ist. Dieses Modell wird bei einem Vergleich mit einer aufgenommenen Szene als Referenz verwendet. Ein 3D-Objektmodell kann zum Beispiel aus SIFT-Merkmalsvektoren bestehen die markante Raumpunkte in einem Objektkoordinatensystem beschreiben. Wenn ein Objekt identifiziert werden soll, dann werden für ein aufgenommenes Bild die SIFT-Merkmalsvektoren bestimmt. Anschließend wird über eine Korrespondenzsuche zwischen Bild- und Raumpunkten ermittelt, ob eine genügende Anzahl von Merkmalsvektoren in einer Ansicht des 3D-Objektmodells übereinstimmen. Nach einer erfolgreichen Identifikation kann mit Hilfe von Gleichung 4 die Pose der Kamera ausgehend vom Ursprung des Koordinatensystems des Referenzmodells bestimmt werden. Die Voraussetzung dafür ist, dass die inneren Kameraparameter bekannt sind. Diese können zum Beispiel mit dem Programm PhotoModeler bestimmt werden. Die Pose eines erkannten Gegenstandes erschließt sich aus der berechneten Kamerapose. Die Berechnung der Kamerapose wird in MEISEL (2008) näher erläutert.

3 Verfahren für 3D-Objekterkennung

3.1 Auszug aus dem Stand der Forschung

3.1.1 Erkennung von 3D-Objekten auf Basis einer Rundumansicht

In ROTHGANGER ET AL. (2006) wird ein Verfahren zur 3D-Objekterkennung vorgestellt, welches auf der Rundumansicht eines zu identifizierenden Objekts basiert. Um die Referenzdaten für einen zu erkennenden Gegenstand in einer Bibliothek von Referenzmodellen vollautomatisch aufzunehmen, werden zu Beginn in der Regel 20 Bilder von dem Objekt erstellt. Dabei werden auf einer ringförmigen Bahn in mittlerer Höhe des stehenden Objekts 16 Bilder aufgenommen. Weiterhin werden 4 Bilder auf einer ringförmigen Bahn über dem Objekt gespeichert. Anschließend werden die markanten Bildpunkte für jedes Bild bestimmt und durch SIFT-Merkmalsvektoren beschrieben. Damit in einem Bild nur markante Bildpunkte des Gegenstands vorhanden sind, werden die Objekte zum Beispiel vor einem weißen Hintergrund aufgenommen der keine Texturen besitzt. Um ein 3D-Objektmodell zu erstellen, wird in überlappenden Bildern ein Vergleich der Merkmalsvektoren vorgenommen. Auf dieser Basis entsteht nach weiteren Berechnungsschritten für jeweils 2 überlappende Bilder ein Teilmodell aus den korrespondierenden Merkmalsvektoren. Diese Teilmodelle werden in weiteren Ausführungsschritten zu einem vollständigen 3D-Objektmodell vereint. Ein solches Objektmodell besteht aus SIFT-Merkmalsvektoren die markante Raumpunkte in einem Objektkoordinatensystem beschreiben.

Die Identifizierung eines Gegenstandes in einem Bild basiert auf dem Vergleich von SIFT-Merkmalsvektoren einer aufgenommenen Szene mit denen des Objektmodells. In dem Verfahren sollen auch Gegenstände wiedererkannt werden, dessen Texturen sich wiederholen. Dies ist eine besondere Anforderung an das Verfahren, weil somit mehrere markante Bildpunkte jeweils mit mehreren markanten Raumpunkten übereinstimmen können. Es muss also ermittelt werden, welche Ansicht des Gegenstandes im Bild vorhanden ist. Die Lösung dieses Problems wird in dieser Ausarbeitung nicht näher beschrieben. Das Verfahren berechnet also, ob sich das zu identifizierende Objekt im Bild befindet und welche Ansicht des Objekts im Bild vorhanden ist. Am Ende dieser Berechnungen ist die Projektionsmatrix bekannt, welche in Abschnitt 2.2.2 beschrieben wird. Die Projektionsmatrix kann dazu verwendet werden, um die Objektpose zu bestimmen.

In ROTHGANGER ET AL. (2006) werden folgende experimentelle Ergebnisse von der Erstellung von 3D-Objektmodellen und der Erkennung von Objekten aufgeführt. Die Erstellung eines 3D-Objektmodells aus 20 Bildern dauerte, unter Verwendung von C++ Programmen auf einem PC mit einer Rechenleistung von 3Ghz, ungefähr 31,5 Stunden. Es wird beschrieben, dass die Erstellung eines solchen Modells auch in ungefähr 3,5 Stunden möglich ist. Allerdings sind dann 4% weniger markante Raumpunkte vorhanden. Eine Steigerung der Geschwindigkeit bringt somit eine Minderung der Qualität mit sich, da die Möglichkeit der Wiedererkennung des modellierten Objekts eingeschränkt wird. Der Vergleich zwischen einem Bild und einem Referenzmodell dauert durchschnittlich 6,7 Minuten. In einer Testreihe wurde eine Wiedererkennungsrate von 88% ermittelt. Die Dauer eines Vergleichs kann um die Hälfte reduziert werden, wobei sich die Wiedererkennungsrate nur marginal verringert. Eine weitere Verringerung der Dauer führt allerdings zu Einbußen bei der Wiedererkennungsrate.

3.1.2 Ein schnelles Verfahren zur 3D-Objekterkennung

In REVAUD ET AL. (2007) wird eine Methode zur 3D-Objekterkennung beschrieben, bei der die Erstellung eines Referenzmodells auf Basis von 2 Bildern geschieht. Die Bilder entstehen wie in Abschnitt 3.1.1 auf einer ringförmigen Bahn um einen Gegenstand. Allerdings werden in diesem Fall nur 2 Bilder aufgenommen, die in einem Winkel von ungefähr 30 Grad zueinander versetzt sind. Somit ist ein Teil eines zu modellierenden Objekts in beiden Bildern sichtbar. Aus diesem Grund können Objekte durch das hier vorgestellte Verfahren in Bildern nicht aus jeder Perspektive identifiziert werden. Der wiedererkennbare Bereich beschränkt sich auf Ansichten, die den Teil des Objekts beinhalten, der auch in den beiden Modellbildern vorhanden ist. Nach der Aufnahme der beiden Modellbilder werden jeweils SURF-Merkmalsvektoren berechnet.

Das SURF-Verfahren basiert laut BAY ET AL. (2006) auf dem gleichen Prinzip wie das SIFT-Verfahren. SURF-Merkmalsvektoren beschreiben markante Bildpunkte mit den Eigenschaften aus Gliederungspunkt 2.1. Es hat sich herausgestellt das die Qualität der Wiedererkennung im Vergleich zu SIFT und GLOH besser ist. In manchen Fällen sogar um bis zu 10%. Weiterhin wurde ermittelt dass SURF hinsichtlich der Bestimmung von Merkmalsvektoren doppelt so schnell ist wie SIFT. SURF-Merkmalsvektoren sind nur halb so groß wie SIFT-Merkmalsvektoren und benötigen somit eine geringere Speichermenge.

Nachdem für beide Bilder SURF-Merkmalsvektoren berechnet wurden, wird eine Korrespondenzsuche durchgeführt. Anschließend werden für beide Bilder nur die markanten Bildpunkte gesichert, die in beiden Bildern vorhanden sind. Die markanten Bildpunkte werden für beide Bilder gesichert, da sie jeweils andere Koordinaten besitzen.

Die Identifikation eines Objekts in einem Bild basiert auf der Theorie, dass die Ansicht des Objekts als Linearkombination von bekannten Ansichten ausgedrückt werden kann. Wenn ein Objekt in einem Bild identifiziert werden soll, dann werden zunächst die markanten Bildpunkte in dem aufgenommenen Bild bestimmt. Daraufhin wird ein Vergleich mit den markanten Bildpunkten des Referenzmodells vorgenommen. Eine bestimmte Anzahl von Korrespondenzen reicht in dem Verfahren allerdings nicht aus. Das kommt daher, dass markante Bildpunkte in einem Bild mit denen des Referenzmodells in hoher Anzahl korrespondieren können, obwohl sich das zu erkennende Objekt nicht im Bild befindet. Ein möglicher Grund dafür wäre, dass die Toleranz für die Abweichung beim Vergleich von markanten Punkten auf einen hohen Wert eingestellt wurde. Nach weiteren Berechnungen wird versucht die Koordinaten der korrespondierenden markanten Bildpunkte des aufgenommenen Bildes jeweils als Linearkombination der korrespondierenden markanten Bildpunkte der beiden bekannten Ansichten des Objektmodells zu beschreiben. Wenn dieser Schritt funktioniert, dann ist das Objekt im Bild vorhanden. Anschließend soll es möglich sein die Objektpose zu schätzen. Dies wird in REVAUD ET AL. (2007) nicht näher beschrieben.

Im Folgenden werden die Testergebnisse für das Verfahren aufgeführt. Für die Experimente wurde ein Athlon mit einer Rechenleistung von 1,85 GHz verwendet. Die Erstellung eines Objektmodells dauerte bei einer Bildgröße von 600 x 500 weniger als 1 Sekunde. Für die Wiedererkennung eines Objekts wurden Bilder mit einer Größe von 800 x 600 und 1600 x 1200 mit einem Referenzmodell verglichen. Für ein Bild mit einer Größe von 800 x 600 dauerte die Wiedererkennung 454 ms. Laut ARTH UND BISCHOF (2008) ist ein Objekterkennungssystem echtzeitfähig,

wenn es mindestens 1 Ergebnis pro Sekunde liefern kann. Für ein Bild der Größe 1600 x 1200 dauerte die Erkennung ungefähr 2 Sekunden.

3.1.3 Regelung auf Basis einer Bildtrajektorie

In HORNUNG UND HEIMANN (2005) wird ein Verfahren beschrieben, dass es ermöglicht ein Objekt in einem Bild zu erkennen und sich dem identifizierten Objekt mit Hilfe einer Regelungskomponente zu nähern. Das Verfahren wurde speziell für Kameras entworfen, bei denen die Kalibrierung ihrer Parameter ein Problem darstellt. Dies ist zum Beispiel der Fall bei Kameras die mit einer Zoomfunktion ausgestattet sind. Die hier vorgestellte Methode führt nicht zu der direkten Bestimmung der Pose eines identifizierten Objekts. Stattdessen nähert sich ein Roboterarm, welcher an seinem Endeffektor über eine Kamera verfügt, einem erkannten Objekt anhand einer Regelung. Bei dem Endeffektor handelt es sich in diesem Fall um eine Hand, die Gegenstände greifen können soll. Im Folgenden wird beschrieben wie Referenzdaten für einen zu identifizierenden Gegenstand aufgenommen werden, wie dieser in einem Bild erkannt wird und wie sich der Roboterarm dem Objekt nähert.

Ein Teil der Referenzdaten besteht aus 64 Bildern die wie in Abschnitt 3.1.1 auf einer ringförmigen Bahn um ein Objekt aufgenommen werden. Anschließend werden für jedes Bild Merkmalsvektoren erstellt, wobei in HORNUNG UND HEIMANN (2005) offen gelassen wird welche Methode für die Beschreibung der markanten Bildpunkte verwendet wird. Die weitere Erstellung von Objektmodellen wird in dieser Ausarbeitung nicht behandelt. Für die folgende Beschreibung reicht die Kenntnis über die 64 Bilder pro Objektmodell aus.

Wenn ein Objekt in einem Bild erkannt werden soll, dann werden zunächst Merkmalsvektoren von der aufgenommenen Szene bestimmt. Daraufhin wird verglichen mit welchem der 64 Bilder des Referenzmodells das aktuelle Bild hinsichtlich der Merkmalsvektoren übereinstimmt. Das korrespondierende Bild des Referenzmodells wird in der weiteren Ausführung Zielbild genannt. Um sich dem identifizierten Objekt zu nähern, wird der Roboterarm über eine Regelung solange bewegt, bis die Koordinaten der markanten Bildpunkte der aufgenommenen Szene mit denen des Zielbildes übereinstimmen. Wenn dies der Fall ist, dann befindet sich der Roboterarm in der gleichen Pose in der das Zielbild aufgenommen wurde. In dieser Pose kann das System in weiteren Schritten ermitteln, wie es einen Gegenstand greifen kann. Damit die Kamera das Zielbild liefert, reichen 4 markante Bildpunkte aus, die zu Beginn erkannt und anschließend verfolgt werden müssen. Auf dem Weg zum Zielbild wird folgender Verarbeitungsschritt in einer Schleife ausgeführt (vgl. Grafik 3.1). Zunächst wird die Abweichung e der Koordinaten s der markanten Bildpunkte der aktuellen Szene von den Koordinaten s_d der markanten Bildpunkte des Zielbildes bestimmt. Der Regler hat als Ausgang die Stellgröße $\dot{q}$, welche die Veränderung der Kameralage angibt.

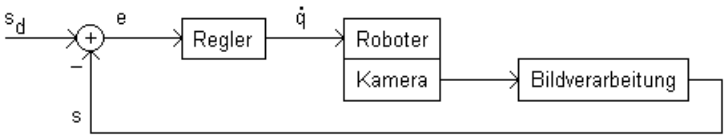

Bild 3.1: Bildbasierte Regelung

Das bisher beschriebene Verfahren wird in der Literatur als IBVS bezeichnet und führt bei einer Kamera, welche nicht kalibriert ist, zu Stabilitätsproblemen. Aus diesem Grund wurde das

Verfahren um eine Bildtrajektorie erweitert. Somit ergibt sich für die Methode die Bezeichnung ITBVS. Um zu einem Zielbild zu gelangen stehen nun mehrere Sub-Zielbilder zur Verfügung, die während der Erstellung eines Referenzmodells aufgenommen wurden. Das erste Sub-Zielbild befindet sich dabei am weitesten vom Zielbild entfernt. Folgende Sub-Zielbilder sind immer etwas dichter an dem Zielbild. Das Verfahren aus Grafik 3.1 wird nun sukzessive für Sub-Zielbilder ausgeführt, bis das eigentliche Zielbild erreicht wird. Das System muss nur kleine Bewegungen durchführen, um zu einem Sub-Zielbild zu gelangen. Das Verfahren wird somit stabil für Kameras, die nicht kalibriert sind.

Laut HORNUNG UND HEIMANN (2005) wurde in einem Experiment ein Gegenstand aus einer Entfernung von ungefähr 1 Meter erkannt und von dem Roboterarm gegriffen. Die Annäherung an das Zielbild dauerte dabei 15 Sekunden, wobei nicht spezifiziert wurde auf welcher Plattform die Berechnungen durchgeführt wurden.

3.2 Bestimmung der Objektpose mit photogrammetrischen Mitteln

Im Folgenden wird eine Idee zur Lösung der Problemstellung aus Kapitel 1 beschrieben, welche an der Fakultät für Technik und Informatik an der HAW Hamburg entwickelt wurde. Für ein zu erkennendes Objekt soll ein 3D-Referenzmodell auf Basis einer Rundumansicht des Objekts erstellt werden. Zunächst soll versucht werden, 3D-Modelle von Gegenständen manuell mit Hilfe des Programms PhotoModeler zu konstruieren. Anschließend soll die Erstellung von 3D-Modellen auf Grundlage der verwendeten Konzepte von PhotoModeler automatisiert werden. Für einen zu identifizierenden Gegenstand soll ein 3D-Modell entstehen, welches aus markanten Raumpunkten in einem Objektkoordinatensystem besteht. Markante Raumpunkte sollen mit Merkmalsvektoren beschrieben werden, wobei noch nicht spezifiziert ist, welches Verfahren verwendet werden soll. Potentielle Kandidaten sind allerdings das SIFT- und das SURF-Verfahren.

Für die Identifikation eines Objekts in einem Bild sollen zunächst die markanten Bildpunkte bestimmt werden. Dafür muss die gleiche Vorgehensweise verwendet werden, wie bei der Beschreibung der markanten Raumpunkte von Objektmodellen. Daraufhin soll eine Korrespondenzsuche zwischen den markanten Bildpunkten und den markanten Raumpunkten des Referenzmodells stattfinden. Wenn eine bisher noch nicht festgelegte Anzahl von Übereinstimmungen vorhanden ist, dann wird angenommen dass sich das Objekt im Bild befindet. Im nächsten Schritt soll die Kamerapose bestimmt werden. Dazu soll die in Gliederungspunkt 2.2.3 beschriebene Vorgehensweise verwendet werden. Für die Bestimmung der Kamerapose gibt es bereits ein funktionierendes Modul, welches im Rahmen einer Masterarbeit an der HAW Hamburg entwickelt wurde. Für das erläuterte Vorgehen soll eine Kamera verwendet werden, deren innere Parameter sich während des gesamten Vorgangs der Erkennung und der Ermittlung der Kamerapose nicht ändern. Dies ist, wie in Abschnitt 2.2.3 erwähnt, die Voraussetzung für die Kalibrierung der externen Kameraparamter. Die Pose eines identifizierten Objekts erschließt sich aus der berechneten Kamerapose.

4 Fazit

Im Folgenden werden die Eigenschaften der Verfahren aus Kapitel 3 zusammengefasst, um anschließend entscheiden zu können welche Vorgehensweise für die Lösung der Aufgabenstel-

lung aus Abschnitt 1 am ehesten geeignet ist. Entscheidend ist die Schnelligkeit und die Qualität der Wiedererkennung eines Objekts in einer aufgenommenen Szene.

Die Methode aus Gliederungspunkt 3.1.1 bietet eine hohe Wiedererkennungsrate. Allerdings werden für einen Identifikationsversuch durchschnittlich 6,7 Minuten benötigt. Aus diesem Grund scheidet das Verfahren aus. Die Vorgehensweise aus Abschnitt 3.1.2 ermöglicht zwar eine schnelle Wiedererkennung, aber ein zu erkennender Gegenstand kann nicht aus jeder Perspektive erkannt werden. Man könnte versuchen, das Verfahren entsprechend anzupassen. Allerdings besteht das Risiko, dass die Erweiterung nicht erreicht werden kann. Somit wird das Verfahren nicht weiter berücksichtigt. In Kapitel 3.1.3 wird eine Methode vorgestellt, die über die Identifikation von Objekten hinausgeht. Das Verfahren beinhaltet bereits eine Annäherung an ein erkanntes Objekt bis zu einer vordefinierten Zielposition, um das Objekt manipulieren zu können. Die Annäherung an ein Objekt in einer Entfernung von 1 Meter dauerte in einem Experiment 15 Sekunden. Die Qualität der Wiedererkennung von Objekten wurde nicht angegeben. Allerdings kann sich der Greifarm einem Gegenstand aus einer beliebigen Perspektive nähern. Da in dem Verfahren Merkmalsvektoren verwendet werden, spricht nichts gegen eine hohe Qualität hinsichtlich der Wiedererkennung von Objekten. Das Verfahren eignet sich somit hervorragend für die hier vorliegende Problemstellung. In Abschnitt 3.2 wird eine vielversprechende Lösungsidee beschrieben. Da es sich bisher nur um einen Lösungsansatz handelt, wird die Idee an dieser Stelle nicht weiter betrachtet.

5 Ausblick

Dieser Abschnitt beschreibt welche Aufgaben gelöst werden müssen, um das gewählte Verfahren aus Abschnitt 3.1.3 verwenden zu können. Weiterhin sollen eventuell vorhandene Risiken identifiziert werden, die bei der Umsetzung auftreten können.

Um die gewählte Vorgehensweise einsetzen zu können, ist es zunächst notwendig, dass man ein tieferes Verständnis über die verwendeten Konzepte erlangt. Dabei handelt es sich hauptsächlich um Konzepte aus dem Bereich der Mathematik und der Informatik. Anschließend können folgende Module nacheinander entwickelt werden.

- Modul zur Erstellung von Referenzmodellen für Gegenstände an die sich ein Greifarm annähern können soll,
- Modul zur bildabhängigen Identifikation von Objekten aus einer beliebigen Ansicht,
- Modul zur Annäherung an einen identifizierten Gegenstand bis zu einer durch das Referenzmodell vorgegebenen Zielposition.

Daraufhin kann das Verfahren mit Hilfe einer Testreihe evaluiert werden. Im Rahmen einer Masterarbeit könnte man versuchen, das Verfahren zu verbessern. Man könnte zum Beispiel probieren, die Annäherung an ein Objekt zu beschleunigen. Das einzige Risiko ist der Zeitfaktor. Unter Umständen reicht die Zeit für die Erstellung der einzelnen Module bis zur Masterarbeit nicht aus. In diesem Fall könnte man in der Masterarbeit ein fehlendes Modul realisieren und versuchen es direkt zu verbessern.

Glossar

GLOH	Gradient Location and Orientation Histogram
HAW	Hochschule für Angewandte Wissenschaften
IBVS	Image Based Visual Servoing
ITBVS	Image Trajectory Based Visual Servoing
Photogrammetrie	Vermessung von Objekten mit optischen Mitteln
SIFT	Scale Invariant Feature Transform
SURF	Speeded Up Robust Features

Literatur

[Arth und Bischof 2008] ARTH, C. ; BISCHOF, H.: Real-time object recognition using local features on a DSP-based embedded system. In: *Journal of Real-Time Image Processing* 3 (2008), Nr. 4, S. 233 – 253

[Bay et al. 2006] BAY, H. ; TUYTELAARS, T. ; GOOL, L. V.: *Speeded Up Robust Features*. 2006. – URL HTTP://WWW.VISION.EE.ETHZ.CH/~SURF/ECCV06.PDF. – Zugriffsdatum: 3.12.2008

[Hartley und Zisserman 2003] HARTLEY, R. ; ZISSERMAN, A.: *Multiple View Geometry in Computer Vision*. 2. Auflage. Cambridge University Press, 2003. – ISBN 0-521-54051-8

[Hornung und Heimann 2005] HORNUNG, O. ; HEIMANN, B.: A model-based approach for visual guided grasping with uncalibrated system components. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2005), S. 226 – 232

[Lange 2008] LANGE, E.: *Verfolgung markanter Raumpunkte in Videobildsequenzen anhand skalierungs- und rotationsinvarianter Merkmale*, Hochschule für Angewandte Wissenschaften Hamburg, Fakultät Technik und Informatik, Department Informatik, Bachelorarbeit, 2008. – URL HTTP://OPUS.HAW-HAMBURG.DE/VOLLTEXTE/2008/611/PDF/BACHELORARBEIT_EMMANUEL_LANGE.PDF. – Zugriffsdatum: 18.11.2008

[Lee et al. 2006] LEE, S. ; KIM, E. ; PARK, Y.: 3D object recognition using multiple features for robotic manipulation. In: *Proceedings of the IEEE International Conference on Robotics and Automation* (2006), S. 3768 – 3774

[Lowe 2004] LOWE, D. G.: *Distinctive image features from scale-invariant keypoints*. 2004. – URL HTTP://WWW.CS.UBC.CA/~LOWE/PAPERS/IJCV04.PDF. – Zugriffsdatum: 18.11.2008

[Meisel 2008] MEISEL, A.: *Skript zu 3D-Bildverarbeitung*. 2008. – URL HTTP://WWW.INFORMATIK.HAW-HAMBURG.DE/UPLOADS/MEDIA/AW_3DBV_V02.PDF. – Zugriffsdatum: 18.11.2008

[Mikolajczyk und Schmid 2005] MIKOLAJCZYK, K. ; SCHMID, C.: A performance evaluation of local descriptors. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 27 (2005), Nr. 10, S. 1615 – 1630

[Nischwitz et al. 2007] NISCHWITZ, A. ; FISCHER, M. ; HABERÄCKER, P.: *Computergrafik und Bildverarbeitung*. 2. Auflage. Vieweg Verlag, 2007. – ISBN 978-3-8348-0186-9

[Revaud et al. 2007] REVAUD, J. ; ARIKI, Y. ; LAVOUÉ, G. ; BASKURT, A.: Fast and cheap object recognition by linear combination of views. In: *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval* (2007), S. 194 – 201

[Rothganger et al. 2006] ROTHGANGER, F. ; LAZEBNIK, S. ; SCHMID, C. ; PONCE, J.: Object modeling and recognition using local affine-invariant image descriptors and multi-view spatial contraints. In: *International Journal of Computer Vision* 66 (2006), Nr. 3. – URL HTTP://LEAR.INRIALPES.FR/PUBS/2006/RLSP06. – Zugriffsdatum: 18.11.2008