Schnelle Tiefenberechnung aus monokularen Farbbildfolgen durch Faktorisierung*

Rüdiger Beß

Lehrstuhl für Mustererkennung (Informatik 5)
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martensstr. 3
D-91058 Erlangen
telephone: +49 9131 85-7891
fax: +49 9131 303811
email: bess@informatik.uni-erlangen.de

Abstract:

Ein von Tomasi, Poelman und Kanade entwickelter Ansatz vereinfacht die Formberechnung aus Bildfolgen im wesentlichen zur Faktorisierung einer Meßwertmatrix. Diese enthält die Bildkoordinaten aller in einer Bildfolge detektierten Punktmerkmale. Voraussetzung dafür ist, daß die perspektivische Projektion der Punktmerkmale ins Bild durch eine lineare Abbildung angenähert wird.

In diesem Beitrag wird die Anpassung des Verfahrens für Echtzeitbedingungen untersucht. Das Ziel liegt in der Bestimmung der dreidimensionalen Struktur eines Objektes für die Objekterkennung.

Die Laufzeit der Formberechnung wird durch die beschriebenen Änderungen um mehr als eine Größenordnung beschleunigt, der Zeitbedarf beträgt damit 5 – 20 ms pro Bild. Bei Näherung der perspektivischen durch eine paraperspektivische Projektion führt die verringerte Genauigkeit jedoch häufig dazu, daß keine eindeutige Lösung für die Form des Objektes bestimmt werden kann. Wie in diesem Beitrag gezeigt wird, kann in diesem Fall auch keine sinnvolle Näherungslösung berechnet werden. Bei Annahme der Parallelprojektion lassen sich jedoch zuverlässig qualitativ gute Ergebnisse erzielen. Die relative Abweichung der Form liegt hier bei 13 %.

Schlüselwörter: *monokulare Bildfolgen, Form aus Bewegung, 3D–Rekonstruktion*

1 Einleitung

Ziel des in diesem Beitrag beschriebenen Verfahrens ist die Berechnung der Form eines Objektes aus einer monokularen Bildfolge in Echtzeit, einerseits zur Objekterkennung, andererseits als Grundlage für eine Rekonstruktion der Objektoberfläche.

Verfahren zur Berechnung von Form aus Bewegung benötigen keine Information über die Kameraposition während der Aufnahmen. Sie eignen sich daher prinzipiell zur Bestimmung von 3D-Information bei frei bewegter, unkalibrierter Kamera. Da die Zuordnung der Merkmale hier durch Verfolgen berechnet wird, muß der Abstand zwischen zwei Bildern relativ klein sein, was zu Bildfolgen mit typischerweise mehreren 100 Bildern führt. Entscheidend für die Anwendung in Echtzeit ist daher ein Schritthalten des Algorithmus mit der Bildfrequenz, für die Verarbeitung eines Bildes stehen damit 40 ms zur Verfügung.

*Die vorgestellte Arbeit wurde von der DFG im Rahmen des SFB 182 gefördert. Nur der Autor ist verantwortlich für den Inhalt.

Der vorliegende Beitrag baut auf den Verfahren von Poelman, Tomasi und Kanade zur Formgewinnung durch Faktorisierung auf [18, 13, 16]. Diese Algorithmen haben gegenüber anderen Algorithmen zur Formberechnung aus Bewegung [2, 3, 9] den Vorteil, daß sie ohne Berechnung der Tiefe als Zwischenwert auskommen und daher sehr robust gegenüber Störungen sind.

Im folgenden werden die Verfahren zur Formberechnung durch Faktorisierung kurz umrissen, eine ausführliche Beschreibung findet sich in der Originalliteratur. Daran anschließend werden die durchgeführten Änderungen erläutert, die zu einer Verringerung der Laufzeit von 3–5 Sekunden auf 5–20 Millisekunden pro Bild führen. Auftretende Stabilitätsprobleme werden diskutiert, schließlich werden Ergebnisse für verschiedene Segmentierungsalgorithmen vorgelegt.

Die Realisierung der Algorithmen erfolgte unter Benutzung von HIPPOS [12, 11], einer NIHCL basierten objektorientierten Klassenbibliothek für die Bildanalyse.

2 Form durch Faktorisierung nach Tomasi, Poelman und Kanade

Die hier umrissenen Algorithmen zur Formberechnung durch Faktorisierung basieren jeweils auf einer linearen Näherung der perspektivischen Projektion. Zunächst wird das Verfahren bei orthographischer Projektion erläutert, danach seine Erweiterung auf die paraperspektivische Projektion.

2.1 Faktorisierung bei orthographischer Projektion

Der Ansatz zur Formberechnung durch Faktorisierung bei orthographischer Projektion stammt von von Tomasi und Kanade [18]. Die Berechnung der Objektform wird hier auf die Zerlegung einer *registrierten Meßwertmatrix* in eine *Formmatrix* und eine *Orientierungsmatrix* zurückgeführt. Die $2F \times P$ *Meßwertmatrix* enthält die gemessenen 2D–Positionen von P Punktmerkmalen in einer Folge von F Bildern, wobei die x– und y–Positionswerte desselben 3D–Punktes jeweils in derselben Zeile stehen und die in einem Bild gefundenen Punkte jeweils in derselben Spalte. Die registrierte Meßwertmatrix wird aus der Meßwertmatrix gebildet indem der Schwerpunkt der Punktkoordinaten eines Bildes jeweils in den Ursprung verschoben wird. Die $2F \times 3$ *Orientierungsmatrix* enthält die Parameter der Kameraorientierung für jedes Bild, die $3 \times P$ *Formmatrix* enthält die 3D–Positionen der Punkte relativ zum Schwerpunkt dieser Punkte. Voraussetzung für die Zerlegung ist die Näherung der perspektivischen Projektion durch eine orthographische Projektion. Die Eindeutigkeit dieser Zerlegung wird durch Auswertung einer Reihe von Nebenbedingungen gewährleistet. Der Vorteil dieses Ansatzes liegt in der Robustheit gegenüber Fehlern, da die Form direkt, ohne Berechnung der Tiefe, bestimmt wird. Im folgenden wird das Verfahren kurz formal beschrieben (nach [18]).

Seien $\{(\boldsymbol{u}_{fp}, \boldsymbol{v}_{fp}) \mid f = 1, \cdots, F,\ p = 1, \cdots, P\}$ die Punktkoordinaten von P Punkten in F Bildern. Dann ist die $2F \times P$ Meßwertmatrix $\boldsymbol{W}$ definiert durch:

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_{11} & \cdots & \boldsymbol{u}_{1P} \\ \vdots & & \vdots \\ \boldsymbol{u}_{F1} & \cdots & \boldsymbol{u}_{FP} \end{bmatrix} \qquad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_{11} & \cdots & \boldsymbol{v}_{1P} \\ \vdots & & \vdots \\ \boldsymbol{v}_{F1} & \cdots & \boldsymbol{v}_{FP} \end{bmatrix} \qquad \boldsymbol{W} = \begin{bmatrix} \boldsymbol{U} \\ \hline \boldsymbol{V} \end{bmatrix}$$

Die Umrechnung in die registrierte Meßwertmatrix $\tilde{W}$ erfolgt durch Subtraktion des Mittelwertes einer Zeile von den Einträgen dieser Zeile. Dieser Mittelwert entspricht genau den Koordinaten des Schwerpunktes der Punktkordinaten in einem Bild:

$$\begin{array}{ll} \tilde{\boldsymbol{u}}_{fp} = \boldsymbol{u}_{fp} - a_f & a_f = \frac{1}{P}\sum_{p=1}^{P} \boldsymbol{u}_{fp} \\ \tilde{\boldsymbol{v}}_{fp} = \boldsymbol{v}_{fp} - b_f & b_f = \frac{1}{P}\sum_{p=1}^{P} \boldsymbol{v}_{fp} \end{array} \qquad \tilde{\boldsymbol{W}} = \begin{bmatrix} \tilde{\boldsymbol{U}} \\ \hline \tilde{\boldsymbol{V}} \end{bmatrix}$$

Seien $\boldsymbol{s}_p = (x_p, y_p, z_p)^T, p = 1, \cdots, P$ die Weltkoordinaten der zu den Punktmerkmalen gehörenden 3D–Punkte und $\boldsymbol{i}_f$, $\boldsymbol{j}_f$ die Basisvektoren der Kamerakoordinatensysteme der einzelnen Bilder. Bei orthogonaler Projektion gilt dann:

$$\begin{array}{l} \tilde{\boldsymbol{u}}_{fp} = \boldsymbol{i}_f^T \boldsymbol{s}_p \\ \tilde{\boldsymbol{v}}_{fp} = \boldsymbol{j}_f^T \boldsymbol{s}_p \end{array} \qquad \tilde{\boldsymbol{W}} = \boldsymbol{R}\boldsymbol{S} \qquad \boldsymbol{R} = [\boldsymbol{i}_1 \cdots \boldsymbol{i}_F \; \boldsymbol{j}_1 \cdots \boldsymbol{j}_F]^T \qquad \boldsymbol{S} = [\boldsymbol{s}_1 \cdots \boldsymbol{s}_P]$$

Dabei faßt die Orientierungsmatrix $\boldsymbol{R}$ die Rotationslage der Kamera in jeder Aufnahme zusammen, die Formmatrix $\boldsymbol{S}$ gibt die 3D–Koordinaten der Punktmerkmale wieder, bis auf Verschiebung und Skalierung.

Aus dieser Beziehung läßt sich das *Rangtheorem* ableiten:

Ohne Störungen hat die registrierte Meßwertmatrix $\tilde{\boldsymbol{W}}$ maximal den Rang drei, dies ergibt sich unmittelbar aus der Größe von $\boldsymbol{S}$.

Bei gestörten Meßwerten läßt sich dieses Theorem erweitern zum *Rangtheorem für gestörte Messungen*:

Die bestmögliche Schätzung $\hat{\boldsymbol{R}}$ und $\hat{\boldsymbol{S}}$ für $\boldsymbol{R}$ und $\boldsymbol{S}$ ergibt sich aus den drei größten Eigenwerten von $\tilde{\boldsymbol{W}}$ zusammen mit ihren korrespondierenden rechten und linken Eigenvektoren:

$$\hat{\boldsymbol{W}} = \hat{\boldsymbol{R}}\hat{\boldsymbol{S}} = \boldsymbol{O}_{3l}\boldsymbol{\Sigma}_3\boldsymbol{O}_{3r}$$

Mit $\hat{\boldsymbol{R}} = \boldsymbol{O}_{3l}\boldsymbol{\Sigma}_3^{1/2}$ und $\hat{\boldsymbol{S}} = \boldsymbol{\Sigma}_3^{1/2}\boldsymbol{O}_{3r}$ läßt sich damit eine Zerlegung von $\hat{\boldsymbol{W}}$ definieren, die jedoch nicht eindeutig ist, da für jede reguläre 3×3 Matrix $\boldsymbol{Q}$ gilt:

$$(\hat{\boldsymbol{R}}\boldsymbol{Q})(\boldsymbol{Q}^{-1}\hat{\boldsymbol{S}}) = \hat{\boldsymbol{W}}$$

Ohne Störungen muß jedoch eine lineare Transformation zwischen der tatsächlichen Form $\boldsymbol{S}$ und $\hat{\boldsymbol{S}}$ und zwischen der tatsächlichen Kameraorientierung $\boldsymbol{R}$ und $\hat{\boldsymbol{R}}$ existieren. Störungen können hier ignoriert werden, solange das Verhältnis zwischen dem dritten und vierten Eigenwert groß genug ist, solange also der Betrag der Störungen klein genug gegenüber den Meßwerten ist.

Aus der Orthonormalität der Kamerabasisvektoren ergeben sich Bedingungen, die zur Bestimmung der gesuchten linearen Transformation hinreichend sind. Es muß gelten: $\boldsymbol{i}_f^T\boldsymbol{i}_f = 1, \quad \boldsymbol{j}_f^T\boldsymbol{j}_f = 1$ und $\boldsymbol{i}_f^T\boldsymbol{j}_f = 0$ und damit $\hat{\boldsymbol{i}}_f^T\boldsymbol{Q}\boldsymbol{Q}^T\hat{\boldsymbol{i}}_f = 1, \quad \hat{\boldsymbol{j}}_f^T\boldsymbol{Q}\boldsymbol{Q}^T\hat{\boldsymbol{j}}_f = 1$ und $\hat{\boldsymbol{i}}_f^T\boldsymbol{Q}\boldsymbol{Q}^T\hat{\boldsymbol{j}}_f = 0$ mit $\boldsymbol{R} = \hat{\boldsymbol{R}}\boldsymbol{Q}$ und $\boldsymbol{S} = \boldsymbol{Q}^{-1}\hat{\boldsymbol{S}}$.

Aus den angegebenen Nebenbedingungen läßt sich durch Lösung des überbestimmten, linearen Gleichungssystems die Matrix $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{Q}^T$ bestimmen. Solange $\boldsymbol{A}$ positiv definit ist, also nur positive Eigenwerte ungleich null hat, läßt sich $\boldsymbol{Q}$ mit Hilfe der Jacobi–Transformation $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{\Lambda}\boldsymbol{L}^T$ bestimmen als: $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{\Lambda}^{1/2}$

Die Herleitungen der angeführten Beziehungen sind in [18] nachzulesen.

2.2 Faktorisierung bei paraperspektivischer Projektion

Poehlman und Kanade [13] haben das Verfahren zur Formberechnung von der orthogonalen Projektion auf die paraperspektivische Projektion erweitert. Einen kurzen Überblick darüber gibt dieser Abschnitt.

Die paraperspektivische Projektion [1] ist wie die orthographische Projektion eine lineare Näherung der perspektivischen Projektion. Während bei der orthographischen Projektion die perspektivische Verzerrung eines Objektes vollständig verloren geht, wird bei der paraperspektivischen Projektion nur die scheinbare Größenänderung innerhalb des Objektes vernachlässigt. Dazu werden die Objektpunkte $\boldsymbol{s}_{fp}$ zunächst mittels Parallelprojektion in eine Hilfsebene abgebildet. Diese verläuft durch den Schwerpunkt des Objektes geht und liegt parallel zur Bildebene. Die Projektionsrichtung wird dabei durch die Gerade zwischen dem Brennpunkt und dem Schwerpunkt der abgebildeten Punkte festgelegt. Die Punkte $\boldsymbol{s}'_{fp}$ auf der Hilfsebene werden schließlich perspektivisch in die Bildebene projiziert.

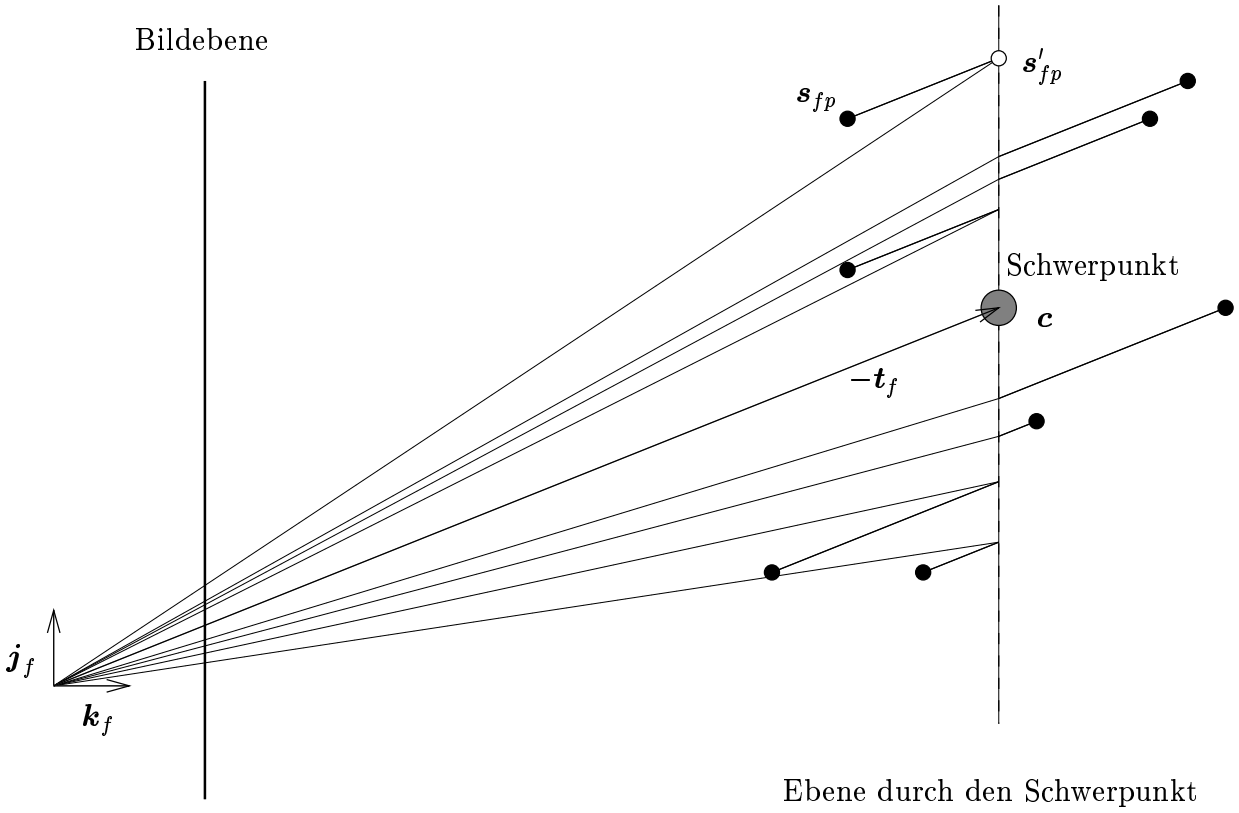

Abb. 1: Paraperspektivische Projektion

Durch die Verringerung des Fehlers in der Approximation der perspektivischen Projektion wird die Genauigkeit der Formberechnung gegenüber dem im letzten Abschnitt beschriebenen Ansatz deutlich erhöht. Allerdings wird das Verfahren empfindlicher gegenüber Störungen.

Das Prinzip des im vorigen Abschnitt beschriebenen Algorithmus bleibt gleich, da nur die Projektionsgleichungen durch eine andere lineare Näherung ersetzt werden. Was sich ändert sind zum einen die Parameter, die die Kameraposition charakterisieren — die Orientierungsmatrix $\boldsymbol{R}$ wird ersetzt durch eine Bewegungsmatrix $\boldsymbol{M}$ — , zum anderen ändert sich die Bestimmung der Matrix $\boldsymbol{Q}$ mit deren Hilfe, wie im vorigen Abschnitt beschrieben, das Ergebnis der Singulärwertzerlegung in die eindeutige Lösung überführt wird.

Seien jetzt $\boldsymbol{i}_f$ und $\boldsymbol{j}_f$ wie vorher die Kamerabasisvektoren, $\boldsymbol{k}_f$ der Normalenvektor in Richtung der optischen Achse, der zusammen mit $\boldsymbol{i}_f$ und $\boldsymbol{j}_f$ ein rechtshändiges Koordinatensystem aufspannt. Sei außerdem $\boldsymbol{c}$ der Schwerpunkt der Objektpunkte und $\boldsymbol{t}_f$ der Brennpunkt, dann gilt für die registrierten Meßwerte:

$$x_{fp} = \boldsymbol{m}_f \boldsymbol{s}_p \qquad y_{fp} = \boldsymbol{n}_f \boldsymbol{s}_p$$

mit

$$z_f = -\boldsymbol{t}_f \boldsymbol{k}_f \qquad c_{x_f} = -\frac{\boldsymbol{t}_f \boldsymbol{i}_f}{z_f} \qquad c_{y_f} = -\frac{\boldsymbol{t}_f \boldsymbol{j}_f}{z_f} \qquad \boldsymbol{m}_f = \frac{\boldsymbol{i}_f - c_{x_f}\boldsymbol{k}_f}{z_f} \qquad \boldsymbol{m}_f = \frac{\boldsymbol{i}_f - c_{x_f}\boldsymbol{k}_f}{z_f}$$

Hier bezeichnet (c_{x_f}, c_{y_f}) den Schwerpunkt der Punktmerkmale eines Bildes. Dieser kann direkt aus dem Bild bestimmt werden. Die Zerlegung der registrierten Meßwertmatrix $\tilde{\boldsymbol{W}} = \boldsymbol{M}\boldsymbol{S}$ erfolgt analog wie im vorigen Abschnitt. Auch hier ergibt sich zunächst keine eindeutige Lösung, gesucht wird die Matrix $\boldsymbol{Q}$ für die gilt:

$$\boldsymbol{M} = \hat{\boldsymbol{M}}\boldsymbol{Q} \quad \text{und} \quad \boldsymbol{S} = \boldsymbol{Q}^{-1}\hat{\boldsymbol{S}}$$

Aus den Abbildungsgleichungen ergeben sich dafür folgende Bedingungen:

$$\frac{|\boldsymbol{m}_f|^2}{1 + c_{x_f}^2} - \frac{|\boldsymbol{n}_f|^2}{1 + c_{y_f}^2} = 0 \quad \text{und} \quad \boldsymbol{m}_f \boldsymbol{n}_f - \frac{c_{x_f} c_{y_f} |\boldsymbol{m}_f|^2}{2(1 + c_{x_f}^2)} - \frac{c_{x_f} c_{y_f} |\boldsymbol{n}_f|^2}{2(1 + c_{y_f}^2)} = 0$$

Um die triviale Lösung $\boldsymbol{M} = \boldsymbol{0}$ (Nullvektor) zu vermeiden, wird $|\boldsymbol{m}_1| = 1$ gesetzt.

Sei $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{Q}^T$, dann wird durch Minimierung des mittleren Fehlerquadrates der $2F$ Bedingungsgleichungen zunächst $\boldsymbol{A}$ bestimmt. Mit $\boldsymbol{M} = \hat{\boldsymbol{M}}\boldsymbol{Q}$ und $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{Q}^T$ eingesetzt in die Bedingungsgleichungen ergibt sich:

$$\frac{\hat{\boldsymbol{m}}_f^T \boldsymbol{A} \hat{\boldsymbol{m}}_f}{1 + c_{x_f}^2} - \frac{\hat{\boldsymbol{n}}_f^T \boldsymbol{A} \hat{\boldsymbol{n}}_f}{1 + c_{y_f}^2} = 0 \quad \text{und} \quad \hat{\boldsymbol{m}}_f^T \boldsymbol{A} \hat{\boldsymbol{n}}_f - \frac{c_{x_f} c_{y_f} \hat{\boldsymbol{m}}_f^T \boldsymbol{A} \hat{\boldsymbol{m}}_f}{2(1 + c_{x_f}^2)} - \frac{c_{x_f} c_{y_f} \hat{\boldsymbol{n}}_f^T \boldsymbol{A} \hat{\boldsymbol{n}}_f}{2(1 + c_{y_f}^2)} = 0$$

Wie bei der orthographischen Projektion läßt sich $\boldsymbol{Q}$, solange $\boldsymbol{A}$ positiv definit ist, mit Hilfe der Jacobi–Transformation $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{\Lambda}\boldsymbol{L}^T$ bestimmen: $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{\Lambda}^{1/2}$

Die Herleitung dieser Beziehungen findet sich in [13].

2.3 Detektion und Ergänzung von Punktmerkmalen

Die Verfahren zur Formberechnung benötigen als Eingabe jeweils Punktmerkmale. Tomasi und Kanade verfolgen dazu die Position von quadratischen Ausschnitten im Bild und nutzen deren Mittelpunkte als Punktmerkmale [17].

Zunächst wird für jedes Pixel ein Ausschnitt mit diesem als Mittelpunkt gebildet. Danach wird mit Hilfe eines Maßes für die Verfolgbarkeit des Ausschnittes jeweils entschieden, welche Ausschnitte weiter betrachtet werden.

Analog dem optischen Fluß [10] wird aus der Grauwertdifferenz zwischen demselben Ausschnitt in zwei aufeinanderfolgenden Bildern die Bewegung dieses Ausschnittes geschätzt. Diese Schätzung wird iterativ verbessert, der Ausschnitt wird um den geschätzten Vektor verschoben und mit demselben Bereich im nächsten Bild verglichen. Aus der neuen Grauwertdifferenz wird eine neue Bewegungsschätzung berechnet. Dies geschieht solange, bis der Unterschied zwischen alter und neuer Bewegungsschätzung unter eine vorgegebene Schranke fällt. Da die Position des verschobenen Ausschnittes im allgemeinen nicht mit dem diskreten Abtastgitter des Bildes zusammenfällt, ist in jedem Iterationsschnitt eine Neuabtastung des Bildausschnittes notwendig.

Dieses Verfahren hat zwei wesentliche Vorteile: Zum einen wird die Position der verfolgten Ausschnitte subpixelgenau erfaßt, zum anderen beschränkt sich die Verfolgung der Punkte auf solche, die sich gut verfolgen lassen.

Ein Nachteil liegt darin, daß die so gewonnenen Merkmale häufig keine charakteristischen Eigenschaften des Objektes repräsentieren, wie zum Beispiel Ecken. Stattdessen werden auch kleinere Unregelmäßigkeiten der Oberfläche erfaßt.

Nach der Merkmalsdetektion müssen fehlende Werte ergänzt werden, da die Vollständigkeit der Meßwertmatrix eine Voraussetzung für die Anwendung der beschriebenen Verfahren zur Formberechnung ist. Für eine Menge von P 3D–Punkten muß in jedem der F Bilder einer Bildfolge die Position bekannt sein. Da diese Forderung in realen Bildfolgen auf Grund von Segmentierungsfehlern und Verdeckungen im allgemeinen nicht erfüllt ist, ist ein Mechanismus notwendig um fehlende Werte zu schätzen.

Tomasi und Kanade lösen dieses Problem, indem sie die Faktorisierung zunächst auf eine vollständige Teilmatrix anwenden, aus dem Ergebnis die fehlenden Werte rekonstruieren und auf diese Weise die Matrix iterativ um jeweils eine Zeile oder Spalte vergrößern [18]. Im störungsfreien Fall reicht es aus wenn von drei Punkten die Koordinaten in vier Bildern bekannt sind und von einem vierten Punkt die Koordinaten in drei dieser vier Bilder. Dann kann die Position des vierten Punktes im vierten Bild rekonstruiert werden (vgl. [19]). Bei realen Bildfolgen muß zur Reduzierung des Einflusses von Störungen möglichst die vollständige Information ausgenutzt werden.

Da zur Bestimmung der fehlenden Punktkoordinaten in jedem Schritt der Iteration eine Singulärwertzerlegung notwendig ist, wird für die Anwendung des Verfahrens in Echtzeit der Zeitaufwand zu groß. Für 250 Merkmale in 400 Bildern sind mehr als 400 Singulärwertzerlegungen erforderlich, mit einer durchschnittlichen Dauer von 5 Sekunden.

3 Segmentierung

Da die berechnete 3D–Form die Grundlage für eine Klassifikation bilden soll, sollten die berechneten 3D–Punkte möglichst charakteristische Merkmale des Objektes wiedergeben. Diese Forderung ist bei der Verfolgung von Bildausschnitten nur bedingt gegeben. Zwar werden auch Merkmale erfaßt, die geometrische Eigenschaften des Objektes wiedergeben, die Detektion läßt sich jedoch nur schwer darauf beschränken [8].

Daher wurden im Rahmen dieser Arbeit zwei weitere Verfahren zur Bestimmung von Punktmerkmalen untersucht: zum einen ein kantenbasiertes Verfahren zur Eckendetektion [5], zum anderen der Monotonieoperator [21, 15, 7]. Der Name des Monotonieoperators leitet sich von seiner Invarianz gegenüber monotonen Grauwerttransformationen des Bildes ab.

Zur Eckendetektion werden zunächst mit einem Nevatia–Babu–Kantenfilter Kantenstärke und Kantenrichtung berechnet, mittels eines Hysterese–Schwellwertverfahrens werden aus den Kanten Kettencodelinien bestimmt. Nach einer Nachberarbeitung zur Schließung von Lücken und Beseitigung von kurzen Linienstücken werden die Ecken basierend auf den Vertices und der Krümmung der Kettencodelinien bestimmt.

Gegenüber der Position der Ecken, die mit per Hand mit Hilfe eines Objektmodells bestimmt werden können, zum Beispiel als Schnittpunkt zweier gerader Linien, ergibt sich eine Abweichung von 1–2 Pixeln.

Der Monotonieoperator beruht auf einer Einteilung jedes Pixels in eine von $n+1$ Klassen $\kappa_i, i = 0, \ldots, n$, wobei n die Anzahl seiner Nachbarn im Bild ist. Jede Klasse κ_i enthält alle Pixel, deren Grauwert größer als der von i Nachbarn ist. Zusätzlich wird eine Abgriffweite L eingeführt, die den Abstand zwischen einem Pixel und den betrachteten Nachbarn bestimmt. Die 4–Nachbarschaft eines Grauwertes $f_{m,n}$ bei Abgriffweite L besteht aus den Punkten $f_{m+L,n}$, $f_{m,n+L}$, $f_{m-L,n}$ und $f_{m,n-L}$. Als Merkmale dienen die Schwerpunkte von Regionen aus Pixeln die der gleichen Klasse zugeordnet wurden. Bei einer genügend hohen Abgriffweite ergeben die Schwerpunkte der Regionen in denen die Pixel der gleichen Klasse κ_0 oder κ_n zugeordnet sind stabile Punktmerkmale.

Zur Verfolgung der Punktmerkmale wird in beiden Fällen jeweils der nächste Nachbar im darauffolgenden Bild bestimmt.

4 Faktorisierung in Echtzeit

Der Zeitbedarf der Formberechnung wird von der Singulärwertzerlegung bestimmt, die Zeit für die Registrierung der Meßwertmatrix und die Auswertung der Bedingungsgleichungen ist dagegen vernachlässigbar.

Die Optimierung der Laufzeit muß sich daher auf eine Verringerung der Anzahl Singulärwertzerlegungen konzentrieren, sowie auf die Beschleunigung oder den Ersatz der einzelnen Singulärwertzerlegung.

4.1 Ergänzung fehlender Punktmerkmale

Wie bereits erwähnt erfordert die iterative Ergänzung der Meßwertmatrix, wie sie von Tomasi und Kanade durchgeführt wird, in jedem Schritt eine Singulärwertzerlegung. Im vorliegenden Beitrag wird die Meßwertmatrix stattdessen in sich überlappende Teilmatrizen aufgespalten. Fehlende Werte werden durch lineare Interpolation der Bildkoordinaten eines Merkmals in vorangegangenen und folgenden Bildern bestimmt. Jede Teilmatrix erlaubt die Berechnung der Form eines Teiles des Objektes der sich zufällig aus den gerade betrachteten Bildausschnitten ergibt. Die überlappenden Werte der Matrizen werden benutzt um die berechneten Teilformen zur gesamten Objektform zusammenzusetzen.

Die Kriterien für die Bildung der Teilmatrizen beinhalten die maximale Anzahl von interpolierten Werten, die maximale Anzahl von aufeinanderfolgenden interpolierten Werten, die Mindestgröße der Teilmatrix und die minimale Anzahl überlappender Zeilen. Die ersten beiden Werte konkurrieren hier mit den letzten beiden Werten. Je weniger Werte in der Teilmatrix interpoliert werden dürfen, desto kleiner wird die vollständige Matrix und desto kleiner wird die Anzahl der gemeinsamen Punkte die in aufeinanderfolgenden Teilmatrizen berechnet werden.

Durch dieses Verfahren reduziert sich die Anzahl der notwendigen Faktorisierungen auf 15–20 pro Bildfolge, durch die geringere Größe der Teilmatrizen sinkt der durchschnittliche Zeitbedarf der Zerlegung auf 0.3 s. Der gesamte Zeitbedarf reduziert sich damit für eine Folge von 400 Bildern auf weniger als 6 Sekunden [20].

4.2 Bestimmung der Eigenwerte

Da die Form– und Bewegungsmatrizen jeweils einen maximalen Rang von drei haben, genügt es die drei betragsmäßig größten Eigenwerte der Meßwertmatrix mit den dazugehörigen Eigenvektoren zu bestimmen. Es ist es demnach nicht notwendig die Singulärwertzerlegung zur Berechnung der Eigenwerte einzusetzen.

Mit dem von Mises – Verfahren [14, 6] steht ein Algorithmus zur Verfügung, der es erlaubt den betragsgrößten Eigenwert einer Matrix $\boldsymbol{B}$ iterativ zu bestimmen, wobei die Anzahl der notwendigen Iterationen vom Quotient zwischen dem betragsgrößten Eigenwert und dem nächsten echt kleineren Eigenwert abhängt. Je kleiner dieser Quotient ist, desto höher ist die Anzahl der notwendigen Iterationen. Dieser Algorithmus konvergiert jedoch nicht, wenn alle Eigenwerte betragsmäßig gleich groß sind, oder wenn der betragsgrößte Eigenwert einer reellen Matrix eine komplexe Zahl ist, oder wenn die Matrix deren Eigenwerte berechnet werden nicht diagonalisierbar ist und der betragsgrößte Eigenwert weniger Eigenvektoren besitzt, als seine algebraische Dimension beträgt. Der erste Fall ist hier unkritisch, wenn alle Eigenwerte betragsmäßig gleich groß sind, läßt sich die Form nicht mehr von den Störungen trennen. In diesem Fall läßt sich die Form auch bei bekannten Eigenwerten nicht bestimmen. Das gleiche gilt für die beiden anderen Einschränkungen. Es genügt daher eine obere Schranke für die zulässige Anzahl an Iterationen zu setzen und den Algorithmus gegebenenfalls abzubrechen.

Ist ein Eigenvektor bekannt, so kann die Deflation eingesetzt werden, um mit Hilfe einer Ähnlichkeitstransformation die Matrix $\boldsymbol{B}$ in eine Matrix $\tilde{\boldsymbol{B}}$ überzuführen. $\tilde{\boldsymbol{B}}$ läßt sich so bestimmen, daß sie die gleichen Eigenwerte wie $\boldsymbol{B}$ hat, bis auf den Eigenwert, der zu dem in der Transformation genutzten Eigenvektor gehört.

Unter erwähnten Voraussetzungen läßt sich die hier beschriebene Kombination aus von Mises – Verfahren und Deflation zur iterativen Berechnung der n betragsgrößten Eigenwerte nutzen. Da die Deflation anfällig gegenüber Rundungsfehlern ist, darf sie jedoch nur dann benutzt werden, wenn nur wenige Eigenwerte bestimmt werden sollen.

In allen bisher getesteten Berechnungen mittels dieses Verfahrens ergaben sich keine wesentlichen Abweichungen der relevanten Eigenwerte vom Ergebnis der Singulärwertzerlegung. [20].

Der Zeitbedarf dieses Verfahrens gegenüber der Singulärwertzerlegung bei der Berechnung der drei größten Eigenwerte im Durchschnitt um den Faktor 5 niedriger, bei sehr großen Matrizen steigt dieser Wert auf über 11 an. Kritisch ist hier die Berechnung des ersten Eigenvektors, da dieser im Einzelfall dem zweiten sehr ähnlich sein kann. Dies kann dazu führen, daß die Berechnung länger dauert, als die Berechnung der Singulärwertzerlegung. Um diesen Fall abzufangen, wird das von Mises–Verfahren nach der Hälfte der Zeit, die eine Singulärwertzerlegung benötigen würde, abgebrochen und die Singulärwertzerlegung durchgeführt. Dies trat bisher bei durchschnittlich 10 Prozent der Zerlegungen auf, der durchschnittliche Zeitgewinn reduziert sich dadurch von 80 Prozent auf 67 Prozent.

4.3 Stabilität des Verfahrens

Während die Zerlegung der Meßdatenmatrix immer durchführbar ist, unabhängig von der Genauigkeit der Meßdaten, gilt dies nicht für die Auswertung der Nebenbedingungen zur eindeutigen Bestimmung der Form. Wie bereits erwähnt sind diese zur Berechnung der Transformationsmatrix $\boldsymbol{Q}$ notwendig, die das Ergebnis der Singulärwertzerlegung $\hat{\boldsymbol{S}}$ in die eindeutige Formmatrix $\boldsymbol{S} = \boldsymbol{Q}^{-1}\hat{\boldsymbol{S}}$ überführt. Dazu wird mit den Nebenbedingungen für die orthographische bzw. die paraperspektivische Projektion jeweils zunächst die Matrix $\boldsymbol{A}$ bestimmt. Nur wenn diese positiv definit ist, also alle ihre Eigenwerte echt größer null, kann $\boldsymbol{Q}$ mit Hilfe der Jacobi–Transformation $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{\Lambda}\boldsymbol{L}^T$ bestimmt werden als $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{\Lambda}^{1/2}$.

Die relativ ungenaue Merkmalsdetektion, die lineare Approximation fehlender Punktmerkmale und die Zerlegung der Meßwertmatrix in kleinere Teilmatrizen erhöhen gegenüber dem ursprünglichen Verfahren den relativen Fehler in den Meßdaten. Eine wesentliche Frage ist daher, ob die Matrix $\boldsymbol{A}$ unter diesen Bedingungen positiv definit ist und was zu tun ist, wenn sie nicht positiv definit ist.

Theoretisch sollten sich für die Annahme einer paraperspektivischen Projektion genauere Werte ergeben, da die Abweichung gegenüber der realen, perspektivischen Projektion geringer ist. Die von Poelman und Kanade [13] angegebenen, veränderten Nebenbedingungen zur Berechnung einer eindeutigen Lösung machen das Verfahren jedoch wesentlich empfindlicher gegenüber Störungen [20]. Die oben erwähnte Matrix $\boldsymbol{A}$ ist bereits bei Rundung idealer Koordinaten auf volle Pixel fast immer nicht positiv definit. Demgemäß konnte bei Annahme der paraperspektivischen Projektion für die im vorliegenden Beitrag getesten realen Bildfolgen mit einem durchschnittlichen Fehler von über einem Pixel keine Form berechnet werden. Bei der Berechnung der Form unter der Annahme einer orthogonalen Projektion treten diese Probleme nicht auf.

Wie hier gezeigt wird ist $\boldsymbol{A}$ entweder positiv definit und es existiert damit eine Lösung, oder es existiert auch keine sinnvolle Näherungslösung. Die beste Näherungslösung der Objektform ist dann maximal zweidimensional.

Gegeben seien n Gleichungen

$$G(\boldsymbol{Q}_i) = \boldsymbol{a}_i\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{b}_i + c_i$$

dann muß die Funktion

$$F(\boldsymbol{Q}) = \sum_{i=1}^{n}(G(\boldsymbol{Q}_i))^2$$

minimiert werden um das beste $\boldsymbol{Q}$ zu bestimmen. Die Lösung ist nicht eindeutig, da für jede orthogonale Matrix $\boldsymbol{D}$ gilt:

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{Q}\boldsymbol{I}\boldsymbol{Q}^T = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{D}^T\boldsymbol{Q}^T$$

und damit $F(\boldsymbol{Q}) = F(\boldsymbol{Q}\boldsymbol{D})$.

Durch eine polare Zerlegung der Matrix $\boldsymbol{Q}$ kann man diese in eine symmetrische positiv semidefinite Matrix $\boldsymbol{B}$ und eine orthogonale Matrix $\boldsymbol{D}$ aufteilen [4, S.614]. Es gibt demnach eine symmetrische, positiv semidefinite Matrix $\boldsymbol{B}$ mit $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{D}$, die die Funktion F minimiert.

Betrachtet man nun die direkte Lösung des Minimierungsproblems mit $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{Q}^T$. Da in diesem Fall die Gleichungen für $\boldsymbol{A}$ linear sind und es keine Einschränkungen durch Nebenbedingungen gibt, existiert genau eine Lösung für $\boldsymbol{A}$.

Sei jetzt $\boldsymbol{B}$ positiv definit, so ist auch $\boldsymbol{B}\boldsymbol{B}^T$ positiv definit und wegen der Eindeutigkeit von $\boldsymbol{A}$ gilt $\boldsymbol{B}\boldsymbol{B}^T = \boldsymbol{A}$. In diesem Fall läßt sich demnach direkt eine eindeutige Lösung bestimmen.

Sei jetzt $\boldsymbol{B}$ nicht positiv definit. Da $\boldsymbol{B}$ positiv semidefinit ist, besitzt $\boldsymbol{B}$ dann mindestens einen Eigenwert gleich Null. In diesem Fall ist $\boldsymbol{Q}$ singulär.

Das Ergebnis der Formberechnung wird dadurch in den zweidimensionalen Raum projiziert [20].

5 Ergebnisse

In den dargestellten Ergebnissen wird eine 3D–Liniendarstellung der Objektes erreicht durch Kombination der Ecken mit den Verbindungslinien der 2D-Segmentierung. Zwei Ecken sind dann mit einer Kante verbunden, wenn in genügend vielen Bildern der Folge eine Linie zwischen diesen Punkten gefunden wird.

Die Bildfolge 'L–Stück' umfaßt 100 Bilder mit 9 verfolgten Punkten. Das vorgestellte Verfahren benötigt hier weniger als 50 ms (0.5 ms pro Bild) für die Berechnung der Form. Auf Grund der geringen Anzahl von Punkten ist das ursprügliche Verfahren nur unwesentlich langsamer. Abb. 2 gibt einen qualitativen Eindruck des Ergebnisses.

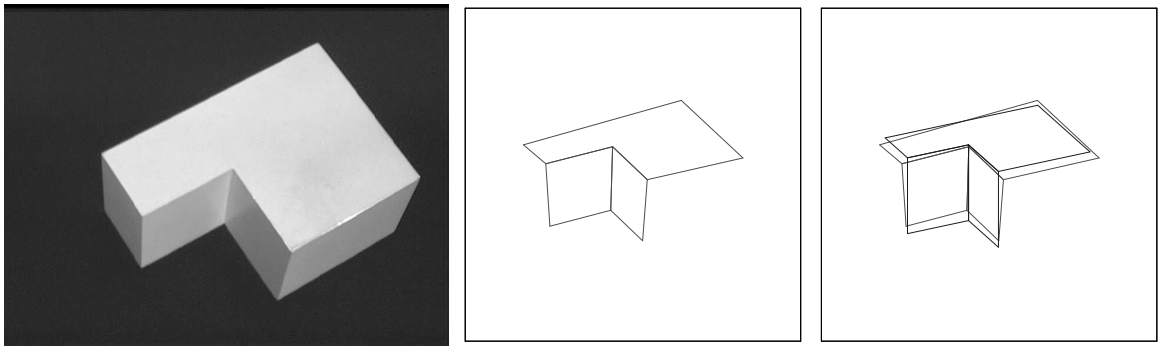

Abb. 2: Ergebnis L–Stück
Aufnahme 60 von 100, Ergebnis der Formberechnung, Ergebnis überlagert mit tatsächlicher Form

Die Bildfolge 'Haus' umfaßt 400 Bilder mit 190 verfolgten Punkten. Abb. 3 zeigt das Ergebnis, verdeckte Linien wurden ausgeblendet. Die Gesamtzeit für die Berechnung der Form beträgt 2.3 Sekunden (5.75 ms pro Bild), davon 0.2 Sekunden für die Aufteilung der Meßwertmatrix, 1.9 Sekunden für die Faktorisierung der Teilmatrizen und 0.2 Sekunden für das Zusammensetzen der Formmatrizen. Nicht betrachtet ist jeweils die Zeit für die Merkmalsdetektion.

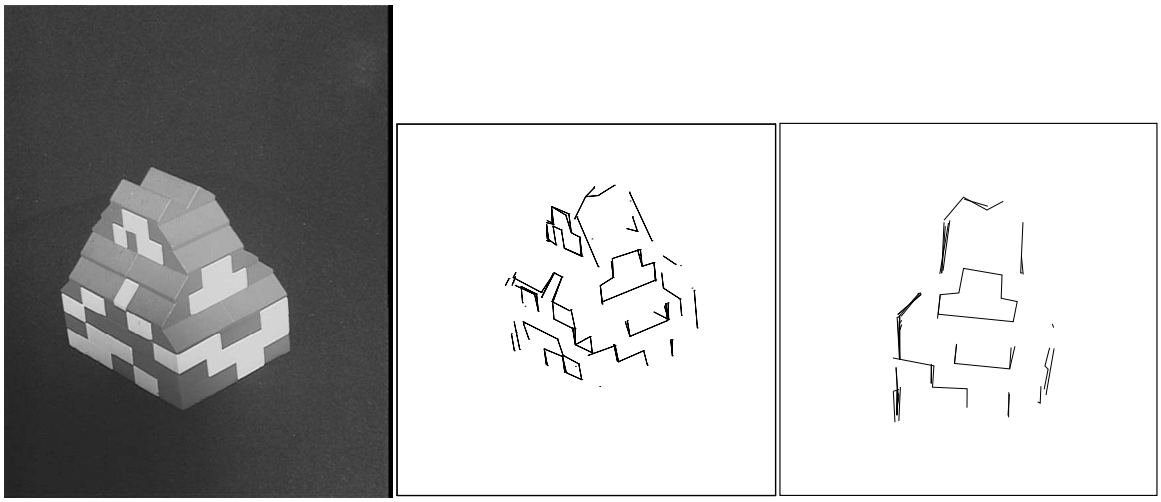

Abb. 3: Ergebnis Haus
Aufnahme 250 von 400, Ergebnis der Formberechnung, Frontansicht des Ergebnisses (verdeckte Linien jeweils ausgeblendet)

Da eine Teilmatrix erst nach vollständiger Kenntnis der zu ihr beitragenden Bilder faktorisiert werden kann, liegt das Ergebnis 0.15–0.4 Sekunden nach Bearbeitung des letzten Bildes vor. Der relative Fehler in der Rekonstruktion der Objekte beträgt im Schnitt 13 %, maximal 18 %.

Die Bildfolge 'Würfel' (Abb. 4) — eine Testfolge für den Monotonieoperator — umfaßt nur 10 Bilder. Hier wurde das Objekt vor der Kamera gedreht, ein Teil der verfolgten Punkte liegt jedoch auf dem unbewegten Hintergrund. Aus diesem Grund und wegen dem geringen Drehwinkel (etwa 5 Grad) ist das Ergebnis etwas verzerrt. Die Bildfolge 'Labor' (Abb. 5), besteht aus 101 Bildern, die Merkmale wurden ebenfalls mit dem Monotonieoperator bestimmt.

Für die Bildfolge 'Würfel' sind die Aufnahmebedingungen unbekannt. In allen anderen Fällen erfolgte die Aufnahme bei völlig unkalibrierter Kamera, die Linsenverzerrung des Objektivs beträgt im Randbereich etwa ein Pixel, die Brennweite der Optik ist 16 mm, die Pixelgröße 0.1×0.1 mm. In den Bildfolgen 'L–Stück' und 'Haus' betrug die Entfernung zum Objekt 500–600 mm, die Größe der Objekte 60–100 mm. Bei der Bildfolge 'Labor' lag die Entfernung bei etwa 1800 mm.

Die Zeitangaben beziehen sich jeweils auf eine HP 735 (99 MHz) mit 124 MIPS.

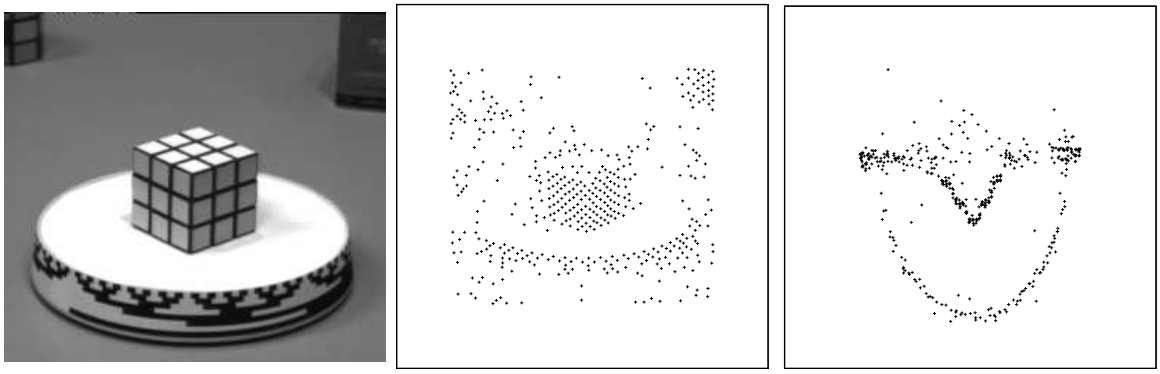

Abb. 4: Ergebnis Würfel
Aufnahme 0 von 10, Ergebnis der Formberechnung, Ansicht des Ergebnisses von oben

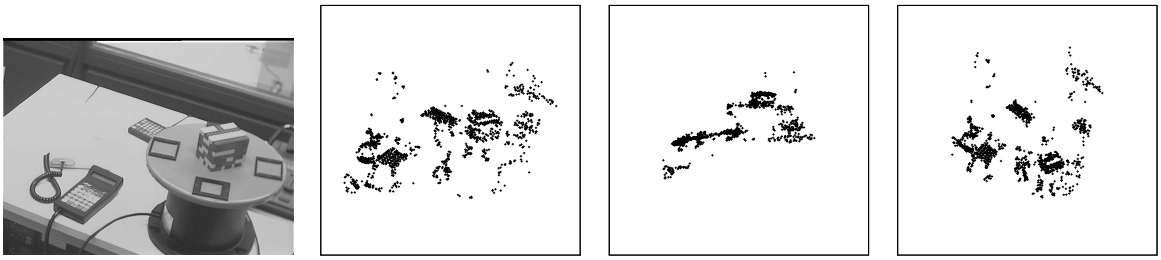

Abb. 5: Ergebnis Labor
Aufnahme 0 von 101, Ergebnis der Formberechnung, Ansicht des Ergebnisses von vorne und von oben

Literatur

[1] John Y. Aloimonos. Perspective approximations. *Image and Vision Computing*, 8(3):177–192, Aug. 1990.

[2] R. C. Bolles, H. H. Baker, and D. H. Marimont. Epipolar–plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1(1):7–55, 1987.

[3] T. Broida, S. Chandrashekhar, and R. Chellappa. Recursive 3D motion estimation from a monocular image sequence. *IEEE trans. on Aerospace and Electronic Systems*, 26(4):639–656, 1990.

[4] D.K. Faddejew and W.N. Faddejewa. *Numerische Methoden der linearen Algebra.* R.Oldenbourg Verlag München Wien, vierte edition, 1976.

[5] M. Harbeck. *Objektorientierte linienbasierte Segmentierung von Bildern* . Dissertation, Technische Fakultät, Universität Erlangen–Nürnberg, Erlangen, 1996.

[6] Eugene Isaacson and Herbert Bishop Keller. *Analyse numerischer Verfahren.* Verlag Harri Deutsch, Zürich und Frankfurt am Main, 1973.

[7] D. Koller. *Detektion, Verfolgung und Klassifikation bewegter Objekte in monokularen Bildfolgen am Beispiel von Straßenverkehrsszenen*, volume 13 of *Dissertationen zur künstlichen Intelligenz.* infix, St. Augustin, 1992.

[8] M. Lechner. Eckendetektion in Grauwertbildern. Studienarbeit, Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), 1995.

[9] L. Matthies, T. Kanade, and R. Szeliski. Kalman filter–based algorithm for estimatin depth from image sequences. *International Journal of Computer Vision*, 3(3):209–236, 1989.

[10] H. Niemann. *Pattern Analysis and Understanding.* Springer, Berlin, 1990.

[11] D. Paulus and J. Hornegger. *Pattern Recognition and Image Processing in C++*. Advanced Studies in Computer Science. Vieweg, Braunschweig, 1995.

[12] D.W.R. Paulus. *Objektorientierte und wissensbasierte Bildverarbeitung*. Vieweg, Braunschweig, 1992.

[13] C. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. Technical Report CMU-CS-92-208, Carnegie Mellon University, Pittsburgh, Pennsylvania, Oct. 1992.

[14] R. Schaback and H. Werner. *Numerische Mathematik*. Springer, Berlin, 1992.

[15] N. Schneider. Merkmalsgestützte Bestimmung von Verschiebungsvektorfeldern. Technical report, Studienarbeit, Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg, Erlangen, 1995.

[16] C. Tomasi and T. Kanade. Shape and motion without depth. In *International Conference on Computer Vision*, pages 91–95, Osaka, Japan, 1990.

[17] C. Tomasi and T. Kanade. Shape and motion from image streams: a factorization method—part 3 **detection and tracking of point features**. Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburgh, Pennsylvania, Apr. 1991.

[18] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, Nov. 1992.

[19] Shimon Ullman. *The Interpretation of Visual Motion*. The MIT Press Series in Artificial Intelligence. MIT Press, Cambridge, Mass., 1979.

[20] M. Vogel. Berechnung von Objektform und Kameraposition aus Bildströmen. Diplomarbeit, Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), 1995.

[21] G. Zimmermann and R. Kories. Eine Familie von nichtlinearen Operatoren zur robusten Auswertung von Bildfolgen. *Proc. of IEEE Workshop on Motion: Representation and Analysis*, pages 96–119, 1986.