Schriftenreihe

Institut für Mess- und Regelungstechnik
Karlsruher Institut für Technologie
Nr. 015

MSc Alexander Bachmann

# Dichte Objektsegmentierung in Stereobildfolgen

Scientific Publishing

Alexander Bachmann

# Dichte Objektsegmentierung in Stereobildfolgen

**Schriftenreihe**
**Institut für Mess- und Regelungstechnik,**
**Karlsruher Institut für Technologie**
Band 015

# Dichte Objektsegmentierung in Stereobildfolgen

von
Alexander Bachmann

Scientific Publishing

Dissertation, Karlsruher Institut für Technologie
Fakultät für Maschinenbau
Tag der mündlichen Prüfung: 14. Mai 2010
Referenten: Prof. Dr.-Ing. C. Stiller, Prof. Dr.-Ing. J. Beyerer

**Impressum**

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe
www.ksp.kit.edu

KIT – Universität des Landes Baden-Württemberg und nationales
Forschungszentrum in der Helmholtz-Gemeinschaft


Diese Veröffentlichung ist im Internet unter folgender Creative Commons-Lizenz
publiziert: http://creativecommons.org/licenses/by-nc-nd/3.0/de/

KIT Scientific Publishing 2010
Print on Demand

ISSN 1613-4214
ISBN 978-3-86644-541-3

# Vorwort

Die vorliegende Dissertation entstand während meiner Tätigkeit am Institut für Mess- und Regelungstechnik des Karlsruher Instituts für Technologie (KIT). Dem Betreuer dieser Arbeit, Herrn Prof. Dr.-Ing. Christoph Stiller, danke ich herzlich für die jederzeit gewährte Unterstützung sowie für die Schaffung einer einzigartigen Arbeitsatmosphäre mit vielen Freiheiten in meiner wissenschaftlichen Arbeit. Für das in mich gesetzte Vertrauen zu Beginn meiner Promotionszeit gilt ihm mein ganz besonderer Dank.

Herrn Prof. Dr.-Ing. Jürgen Beyerer danke ich für die Übernahme des Korreferats und sein Interesse an meiner Arbeit.

Bei meinen Kollegen am MRT bedanke ich mich für zahlreiche anregende Diskussionen, inspirierende Sommerseminare und eine äußerst angenehme Zusammenarbeit. Hervorheben möchte ich hier meinen Mentor Herrn Thao Dang und Herrn Sören Kammel, denen ich herzlich für die mühsame Arbeit des Korrekturlesens, für Motivation und viele wertvolle Anregungen danke. Dem Sekretariat danke ich für die unkomplizierte Hilfe in allerlei administrativen Belangen. Den Werkstätten sowie Herrn Werner Paal schulde ich großen Dank für ihre Unterstützung in allen praktischen Belangen.

Besonders hervorheben möchte ich die Unterstützung dieser Arbeit durch die Deutsche Forschungsgemeinschaft im Rahmen des Sonderforschungsbereichs „Kognitive Automobile". Für die angenehme Zusammenarbeit möchte ich auch der Firma Honda danken. Weiterhin möchte ich der Firma Daimler in Person von Herrn Alexander Barth für die Überlassung umfangreicher Datensätze danken.

Gewidment ist diese Arbeit meiner Familie, insbesondere meiner Frau Sandra und meiner Tochter Emilia, mit denen ich einzigartige Momente erleben darf und ohne deren Unterstützung und Geduld diese Dissertation nicht hätte gelingen können.

Karlsruhe, im Mai 2010 Alexander Bachmann

# Kurzfassung

Bei der Entwicklung neuer und innovativer Funktionen im Bereich des autonomen Fahrens und der Fahrerassistenzsysteme ist die Aufgabe der Umgebungswahrnehmung von fundamentaler Bedeutung. Hierbei nimmt die visuelle Umfelderfassung eine zunehmend wichtige Stellung ein. Der Grund hierfür liegt in dem ausgesprochen hohen Abstraktionspotential der Bilddaten, durch welches eine Umgebung beliebig komplex dargestellt werden kann. Um die Informationsfülle der Bilddaten für eine praktische Anwendung nutzbar zu machen, greifen derzeitige Auswertesysteme hierbei meist nur auf einen sehr eingeschränkten Teil der Daten zu. Die rasante Entwicklung im Bereich der Rechen- und Speichertechnologie erlaubt mittlerweile jedoch den Einsatz anspruchsvoller Modelle und Algorithmen zur Bildanalyse, die eine detaillierte Abbildung der Umgebung im Inneren eines Rechners erlauben. Besonders wünschenswert ist hierbei die Repräsentation der umgebenden Szene in Form einer dichten Segmentierung, welche relevante Objekte vollständig aus der restlichen Umgebung herauslöst und identifiziert. Die in dieser Arbeit vorgestellte Szenensegmentierung liefert eine solche vollständige und minimal einschränkende Beschreibung der dreidimensionalen Szene. Dabei wird ein einzelnes Objekt als eine, relativ zur Kamera, räumlich und zeitlich gleichförmig bewegte Gruppierung von Szenenpunkten beschrieben. Eine Besonderheit des Verfahrens ist die Tatsache, dass die einzelnen Teilaufgaben der dreidimensionalen Rekonstruktion, Bewegungsschätzung und Segmentierung dabei in einem gemeinsamen Modell beschrieben und in verzahnter Reihenfolge gelöst werden. Die einzelnen Modellparameter werden dabei robust durch merkmalsbasierte Verfahren bestimmt, wobei das aktuelle Segmentierergebnis in der Form eines probabilistischen Assoziationsgewichtes mit in den rekursiven Schätzprozess auf der Basis eines Kalman-Glätters integriert wird. Hierdurch kann eine fortlaufende Verbesserung der verkoppelten Schätzgrößen erreicht werden, deren Güte über die Zeit kontinuierlich zunimmt. Die Erwartung der räumlichen und zeitlichen Konsistenz zusammengehöriger Szenenbereiche wird explizit durch das Modell eines Markov-Zufallsfeldes berücksichtigt. Zur Lösung der globalen Optimierungsaufgabe werden aktuelle Graphenschnittverfahren eingesetzt. Das Ergebnis der bewegungsbasierten Szenensegmentierung ist eine Menge von Bildbereichen, die vom Menschen eindeutig als unabhängig bewegte Objekte interpretierbar sind. Die Leistungsfähigkeit des Verfahrens wird anhand von realen und synthetischen Bildsequenzen aufgezeigt. Die Ergebnisse sind hierbei den meisten aktuellen Objektdetektionsverfahren im Bereich der mobilen Umfeldwahrnehmung deutlich überlegen und unterstreichen das hohe Potenzial des Verfahrens.

**Schlagworte:** Objektdetektion – Bewegungsschätzung – Datenassoziation

# Abstract

One of the cornerstones in the development of automotive driver assistance systems is the comprehensive perception and understanding of the environment in the vicinity of the vehicle. In this context, vision sensors provide a rich and versatile source of information. In the past, most of the systems only used a small fraction of this information due to the limited computational power aboard a vehicle. Today this is going to change as modern driver assistance systems are expected to carry out sophisticated tasks, pushing the vehicle more and more into intelligent interaction with the driver. The precondition for such action is a detailed description of the environment. Optimally, such a description reflects the human perception and segregates the scene densely into physically relevant and meaningful regions. In this work, such regions represent other traffic participants. The object detection task is performed based on the relative motion of these objects to the observer. To obtain a dense representation of the observed scene, object detection is formulated as an image segmentation task. Different to previous approaches in three dimensional scene analysis, the involved tasks of scene reconstruction, motion estimation and image segmentation are formulated in a joint model and solved in an iterative way. Within the filter framework the actual segmentation result is integrated into the estimation task by means of a data association process which weights each observation according to a probabilistic measure. The method first estimates the model parameters for a number of object hypotheses based on a set of feature points in a recursive manner. The evolving parameter estimates are then used to recover the scene depth from spatially and temporally separated views. Given the dense depth map and the set of tracked object parameters, scene segmentation is performed. A Markov Random Field is used to express expectations on spatial and temporal continuity of objects. The performance of the approach is evaluated on real and synthetic image data.

**Keywords:** Object Detection – Motion Estimation – Data Association

# Inhaltsverzeichnis

# Symbolverzeichnis

## Abkürzungen

| | |
|---|---|
| 2D/3D | zwei-/dreidimensional |
| MAP | Maximum-A-Posteriori |
| ML | Maximum-Likelihood |
| MRF | Markov-Zufallsfeld (engl. *Markov Random Field*) |
| GC | Graphenschnitt (engl. *Graph Cut*) |
| BP | Belief-Propagation |
| TLS | engl. *Total-Least-Squares* |
| EM | engl. *Expectation-Maximization* |
| KF/EKF | Kalman-Filter/erweitertes Kalman-Filter |

## Notationsvereinbarungen

| | |
|---|---|
| Skalare | nicht fett, kursiv: $x$, $y$, $z$, $X$, $Y$, $Z$, ... |
| Vektoren | fett, nicht kursiv: $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, ... |
| Matrizen | fett, nicht kursiv, groß: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, ... |
| Mengen | kalligraphisch, groß: $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, ... |
| Konstanten, Bezeichner | nicht kursiv: a, b, c, ... |
| Zufallsgrößen | Zufallsvariablen und Zufallsfelder werden mit *Großbuchstaben* bezeichnet, für Realisationen daraus werden die entsprechenden *Kleinbuchstaben* verwendet |

# Operatoren

| | |
|---|---|
| $\arg\{\cdot\}$ | Argument einer Funktion |
| $\mathrm{E}[\cdot]$ | Erwartungswert |
| $\|\cdot\|$ | Euklidische Norm |
| $\exp[\cdot]$ | Exponentialfunktion |
| $\nabla$ | Nabla-Operator |
| $P(X)$ | Wahrscheinlichkeitsfunktion; Wahrscheinlichkeit für das Eintreten des Ereignisses $X$ |

# Indizes

| | |
|---|---|
| $\widehat{\boldsymbol{\chi}}_n^{\mathrm{b}}$ | bewegungsprädizierte Position des $n$-ten Elements einer Menge von $N$ Objektpunkten, $n = (1,\ldots,N)$ |
| $\widehat{\boldsymbol{\chi}}_n^{\mathrm{s}}$ | stereoskopisch versetzte Position des $n$-ten Elements |
| $\boldsymbol{\Theta}_t, \mathrm{g}_t, \ldots$ | Größen zum Zeitschritt $t$, $t = (0,\ldots,T)$ |
| $\boldsymbol{\Theta}^k, \mathbf{I}^k, \ldots$ | Größen zum Iterationszeitpunkt $k$, $k = (0,\ldots,K)$ |
| $\boldsymbol{\theta}^j, \mathcal{X}^j, \ldots$ | Größen der Objekthypothese $j$, $j = (1,\ldots,J)$ |
| $\mathrm{g}^{\mathrm{l}}, \mathrm{g}^{\mathrm{r}}$ | Bilder der linken und rechten Kamera |

# Häufig verwendete Symbole

| | |
|---|---|
| $\mapsto$ | Abbildung |
| $:=$ | Definition |
| $\cong$ | projektive Identität |
| $\propto$ | Proportionalität |
| $\hat{\mathbf{x}}$ | Schätzwert von $\mathbf{x}$ |
| $\mathbf{x}^{\mathrm{T}}$ | Transposition des Vektors $\mathbf{x}$ |
| | |
| $c_u$ | Clique; Teil der Menge $\mathcal{C}_u$ |
| $\mathcal{C}$ | Menge aller Cliquen, bestehend aus Teilmengen $\mathcal{C}_u$ |
| $c(\mathcal{Q},\mathcal{S})$ | Schnitt durch Graphen G; Zerlegt Knotenmenge $\mathcal{P}$ in disjunkte Mengen $\mathcal{Q}$ und $\mathcal{S}$ |
| $C$ | Gütemaß |

| | |
|---|---|
| $\mathbf{d} = (d_x, d_y)^{\mathrm{T}}$ | diskrete Bildverschiebung; Element des Verschiebungsfeldes $\mathcal{D} = \{\mathbf{d}_n \vert n \in \mathcal{R}\}$ |
| $\mathrm{D}(\cdot)$ | Datenterm |
| $\mathbf{D}(\cdot)$ | Vektorfunktion der logarithmierten Likelihoodfunktion |
| $d$ | Distanzmaß zweier Cluster |
| $\mathbf{E}$ | Restfehlerkarte, mit Komponenten $\varepsilon \in \mathbf{E}$ |
| $\tilde{\mathbf{E}}$ | Vollständige Daten, bestehend aus Restfehlerkarte $\mathbf{E}$ und unbekannter bzw. verdeckter Labelvariable $\mathbf{l}$ |
| $\mathbf{e}$ | Einheits-/Basisvektor |
| $\mathcal{F}$ | Diskreter Konfigurationsraum |
| $f(\mathcal{Q}, \mathcal{S})$ | Fluss durch den geschnittenen Graphen G |
| $\mathcal{G}_T$ | Stereobildsequenz der Länge $T$, bestehend aus einer Menge von Bildern $\{\mathrm{g}(\mathcal{X}, t) \vert t \in \mathcal{T} = \{0, \ldots, T\}\}$ |
| $\mathrm{g}(\mathcal{X}, t)$, $\mathrm{g}_t(\mathcal{X})$, $\mathrm{g}_t$ | Einzelbild der Bildsequenz zum Zeitschritt $t$ |
| $\mathrm{g}_t(\mathbf{x}_n)$, $\mathrm{g}_{n,t}$ | Grauwert des $n$-ten Elements einer Menge von $N$ Bildpunkten |
| $\mathrm{G}(\mathcal{P}, \mathcal{E}, \mathcal{W})$ | Graph, bestehend aus einer Menge von Knoten $\mathcal{P}$, Kanten $\mathcal{E}$ und Kantengewichten $\mathcal{W}$ |
| $\mathbf{G}$ | Glätter-Verstärkungsmatrix |
| $H$ | Hamiltonfunktion |
| $\mathbf{H}_{\boldsymbol{\theta}}$ | Jakobimatrix der Beobachtungsgleichung |
| $\mathcal{I}$ | Indexmenge |
| $\mathcal{J}$ | Menge der Objekthypothesen |
| $\mathbf{K}$ | Kalman-Verstärkungsmatrix |
| $\mathcal{L}$ | Menge möglicher Labelwerte |
| $\mathbf{l}(\mathcal{X}, t), \mathbf{l}_t$ | Segmentierung |
| $\mathcal{N}_n$ | Nachbarschaftssystem von Element $n$ |
| $\mathbf{n} = (a, b, c)^{\mathrm{T}}$ | Parametervektor einer Ebene |
| $\mathcal{O}$ | Cluster der Menge $\mathcal{M}$ |
| $\mathbf{P}$ | Kovarianzmatrix des Zustandes $\boldsymbol{\theta}$ |
| $q, s$ | Abschlussknoten des Graphen G |
| $\mathrm{Q}(\cdot \vert \cdot)$ | Q-Funktion des EM-Verfahrens |
| $\mathcal{R}$ | Indexmenge des Bildrasters |
| $\mathbf{R}, \mathbf{T}$ | Extrinsische Parameter des Stereokamerasystems |
| $\mathbf{u} = (u, v)^{\mathrm{T}}$ | Element des optischen Flussfeldes $\mathcal{U} = \{\mathbf{u}_n \vert n \in \mathcal{R}\}$; Approximation der wahren Bildbewegung $\dot{\mathbf{x}} = (\dot{x}, \dot{y})^{\mathrm{T}} \in \{\dot{\mathbf{x}}_n \vert n \in \mathcal{R}\}$ |

| | |
|---|---|
| $V_c$ | Cliquenpotential |
| $\mathbf{V}(\cdot)$ | Vektorfunktion der logarithmierten Prioriverteilung |
| $\mathbf{X}_{\mathcal{I}}$ | Zufallsprozess |
| $\mathbf{x}_{\mathcal{I}}$ | Realisation des Zufallsprozess |
| $\mathbf{x} = (x,y)^{\mathrm{T}}$ | 2D-Ortskoordinaten im Bildkoordinatensystem |
| $\mathsf{x} = (x,y,1)^{\mathrm{T}}$ | 2D-Ortskoordinaten in homogenen Koordinaten |
| $\mathbf{X} = (X,Y,Z)^{\mathrm{T}}$ | 3D-Ortskoordinaten im Kamerakoordinatensystem |
| $\mathsf{X} = (X,Y,Z,1)^{\mathrm{T}}$ | 3D-Ortskoordinaten in homogenen Koordinaten |
| $\mathbf{x}^{\mathrm{m}}$ | Merkmalspunkt |
| $\mathcal{X}^j$ | Bildsegment, welches Objekthypothese $j \in \mathcal{J}$ zugewiesen ist |
| $\boldsymbol{Z}$ | Tiefenkarte; Realisation aus diskretem Strukturraum $\mathcal{Z}$ |
| $\mathsf{Z}$ | Zustandssumme |
| $\mathsf{z}$ | Zeithorizont des Kalman-Glätters |
| $\Delta$ | Disparitätskarte; Realisation aus diskretem Disparitätsraum $\mathcal{P}$. |
| $\boldsymbol{\theta} = (\mathbf{v}, \boldsymbol{\xi}, \boldsymbol{\phi})^{\mathrm{T}}$ | Parametervektor des Objektmodells, bestehend aus dem Bewegungsvektor $\mathbf{v} = (\omega_x, \omega_y, \omega_z, t_x, t_y, t_z)$, dem Lageparametersatz $\boldsymbol{\xi} = (\mathbf{M}, \boldsymbol{\Sigma})$ mit Position $\mathbf{M}$ und Ausdehnung $\boldsymbol{\Sigma}$ im 3D Raum und dem Parametervektor $\boldsymbol{\phi}$ des Verteilungsmodells |
| $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{Z}\}$ | Parametersatz des Szenenmodells |
| $\lambda$ | Regularisierungskonstante |
| $\Lambda$ | A-Priori Wahrscheinlichkeit einer Objekthypothese |
| $\Pi$ | Projektion in euklidischen Koordinaten |
| $\Pi^{-1}$ | Inverse Projektion in euklidischen Koordinaten |
| $\pi$ | Bedingte Wahrscheinlichkeitsverteilung der Segmentierung |
| $\hat{\boldsymbol{\varsigma}}$ | Beobachtungsvektor eines Merkmals |
| $\Sigma_{\mathbf{ee}}, \Sigma^*_{\mathbf{ee}}$ | Beobachtungsrauschen der Merkmale mit Systemrauschen $\mathbf{e}$; $*$ kennzeichnet das mit $\boldsymbol{\tau}$ gewichtete Beobachtungsrauschen |
| $\boldsymbol{\tau}$ | Hypothesenwahrscheinlichkeit des Merkmals |
| $\boldsymbol{\varphi}$ | Modellparameter des Zufallsfeldes |
| $\widehat{\boldsymbol{\chi}}_{\mathrm{x}}$ | 3D-Ortskoordinaten des korrespondierenden Punktes zu $\mathbf{X}$ |
| $\widehat{\boldsymbol{\chi}}$ | 2D-Ortskoordinaten des korrespondierenden Punktes zu $\mathbf{x}$ |
| $\omega \in \Omega_{\mathcal{I}}$ | Stichprobenraum einer einzelnen Zufallsvariable, mit $\Omega_{\mathcal{I}}$ als Stichprobenraum des gesamten Zufallsfeldes |
| $\boldsymbol{\omega}, \mathbf{t}$ | Rotations- und Translationsparameter des Modells der Starrkörperbewegung |

KAPITEL 1

# Einleitung

Diese Arbeit befasst sich mit der bildbasierten Analyse einer Verkehrsszene für eine Anwendung im Bereich kognitiver Automobile und Fahrerassistenzsysteme. Eine Szene[1] ist hierbei definiert als eine Anordnung von Objekten, die einen räumlich-zeitlichen Ausschnitt der realen Welt beschreibt. Bezüglich der Wahrnehmung haben bildgebende Sensoren hier ein besonders hohes Potenzial, da es durch das passive Messprinzip keine gesetzlichen Einschränkungen hinsichtlich der Zulassung im öffentlichen Straßenverkehr gibt. Weiterhin sind die Infrastruktur und das Verkehrsgeschehen stark auf visuelle Wahrnehmung ausgerichtet und somit nur bildgebend voll zu erfassen. Damit verbunden weist der Informationsgehalt des Sensorsignals ein ungleich höheres Abstraktionspotential auf als herkömmliche Umfeldsensoren. Das Ziel der Bildauswertung ist die vollständige Unterteilung der Szene in Bereiche, die für eine Bewertung von Verkehrssituationen oder auch für das aktive Eingreifen in das Verkehrsgeschehen von besonderem Interesse sind. Denkbar sind hier z. B. Systeme, die den Fahrer mit sicherheitsrelevanter Information versorgen, kritische Situationen durch Brems- oder Lenkeingriffe auflösen oder auch ganze Fahrmanöver planen. Offensichtlich spielen Verkehrsteilnehmer die sich selbst bewegen, hierbei eine ausgesprochen wichtige Rolle. Die vollständige Herauslösung und Identifizierung dieser Objekte aus der restlichen Umgebung, entsprechend der menschlichen Wahrnehmung, stellt dabei das optimale Ergebnis dar. Bei der Repräsentation der Szene und darin enthaltener Objekte liegt die große Herausforderung hierbei in der Komplexität und Vielfalt realer Verkehrssituationen. Als Konsequenz existieren unterschiedlichste Möglichkeiten zur Darstellung und Beschreibung. Eine nützliche Taxonomie in diesem Zusammenhang liefert [Marr, 1982] mit seinem hierarchischen Interpretationsmodell. Die einzelnen Ebenen dieses Modells werden durch aufgabenspezifische Teilprozesse erzeugt, wie in Abbildung 1.1 skizziert. Alle Verfahren der bildbasierten Szenenanalyse lassen sich einem dieser Prozesse zuordnen.

[1] Vom griechischen „Skene" für Bühne.

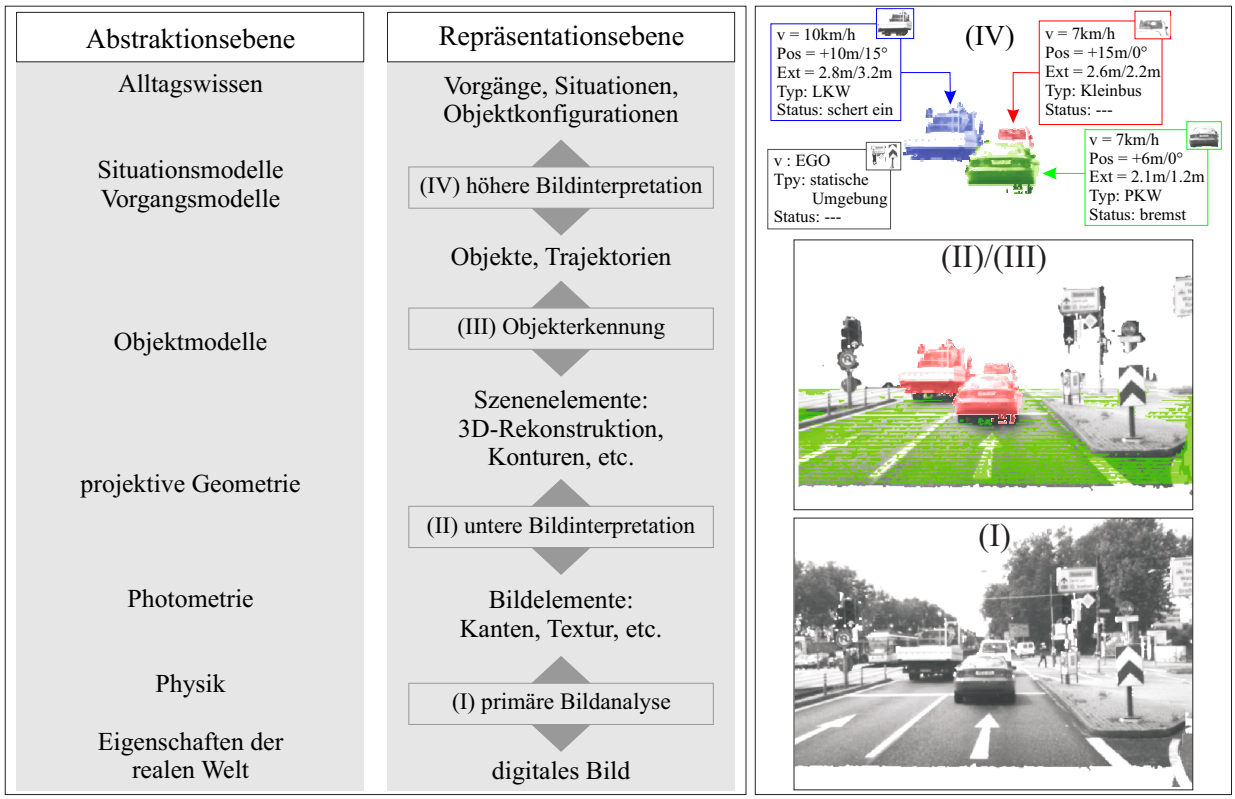



**Abbildung 1.1**: Links: Szeneninterpretation als hierarchischer Prozess: Die Abstraktionsebene erlaubt Rückschlüsse vom Bild auf die Szene. Die unterschiedlichen Repräsentationsebenen werden hierbei durch die entsprechenden Teilprozesse (I)-(IV) erzeugt. Rechts: Einordnung der einzelnen Verarbeitungsschritte des hier vorgestellten Verfahrens in die Taxonomie von [Marr, 1982] am Beispiel einer Verkehrsszene mit bewegten Objekten.

Die primäre Bildanalyse (I) wertet auf der untersten Verarbeitungsebene die radiometrischen Eigenschaften jedes Bildpunktes in der zweidimensionalen (2D) Bildebene aus und gelangt zu einer rudimentären Beschreibung der Szene in Form von z. B. Kanten, Ecken oder der Bildtextur. Auf dieser Ebene werden die für diese Arbeit benötigten Merkmale und Deskriptoren aus dem Bild extrahiert.

Bei der unteren Bildinterpretation (II) werden Bildelemente als Szenenelemente beschrieben und als Abbildung von Teilen einer dreidimensionalen (3D) Szene interpretiert. Auf dieser Stufe soll eine zentrale Aufgabe des Bildverstehens gelöst werden: die Extraktion von Realwelteigenschaften aus Bildeigenschaften. Dazu zählt insbesondere die Rekonstruktion der dreidimensionalen Szene. Es werden außerdem einfache Modellannahmen bezüglich der Szenengeometrie eingeführt. Ein solches Geometriemodell wird in dieser Arbeit genutzt, um eine Vorsegmentierung der Szene durchzuführen. Im mittleren Bild auf der rechten Seite von Abbildung 1.1 ist das Modell in grün eingeblendet. Die Vorsegmentierung hat zum

Ziel, geometrisch nicht plausible Szenenbereiche in höheren Ebenen der Szeneninterpretation auszublenden.

Bei der Objekterkennung (III) werden schließlich semantisch fassbare Objekte in den bisher extrahierten Bilddaten identifiziert und aus den nicht relevanten Daten herausgelöst[2]. Die Mehrheit der Verfahren beschränkt sich dabei auf einzelne Objekttypen, die für eine spezifische Anwendung von Interesse sind. Hierbei können die Ansätze prinzipiell in zwei Gruppen eingeteilt werden: modellgetriebene und datengetriebene Ansätze.

*Modellgetriebene Ansätze* bauen auf meist komplexen Objektmodellen auf, deren Charakteristik vorher bekannt sein muss. Es werden vordefinierte, oft parametrisierte Modelle eines bestimmten Objekttyps verwendet. Je nach Aufgabenstellung und Szenencharakteristik können Objekthypothesen z. B. anhand der spezifischen Geometrie, deren Farbverteilung oder bestimmter Modellprimitive generiert werden [Bertozzi u. a., 1997; Kalinke u. Tzokamkas, 1998]. Diese Modelle oder Schablonen werden dann mit den Bilddaten verglichen und auf ihre Ähnlichkeit hin überprüft. Modellbasierte Verfahren haben sich in einfachen Verkehrsszenarien, in denen die Vielfalt der Objekte und der Betrachtungswinkel der Kamera eingeschränkt ist, vielfach bewährt. Für Szenarien, in denen das Erscheinungsbild von Objekten stark variiert, sind jedoch aufwändige Modelle notwendig, die eine praktische Anwendung stark einschränken. Alternativ dazu kann die Charakteristik eines Objekttyps auf der Basis eines Trainingsdatensatzes erfasst werden. Dieser Datensatz beinhaltet unterschiedliche Ausprägungen eines einzelnen Objekttyps. In einem Trainingsschritt werden aus den Trainingsbildern charakteristische Deskriptoren erzeugt. Eine Sammlung von solchen Deskriptoren kann dann zu einer Gesamtbeschreibung des Objekttyps zusammengefasst werden. Durch den Einsatz effizienter Klassifikationsverfahren können die aus dem Bild extrahierten Deskriptoren mit den trainierten Objektbeschreibungen abgeglichen und auf die mögliche Existenz einer bestimmten Klasse geprüft werden [Papageorgiou u. Poggio, 2000; Mohan u. a., 2001; Schwab, 2006; Balthasar, 2007; Bachmann u. Dang, 2008; Bachmann u. Balthasar, 2008; Bachmann u. Lulcheva, 2009a]. Der wesentliche Nachteil der Verfahren ist das aufwändige Anlegen einer repräsentativen Datenbank für jeden Objekttyp und das teilweise zeitintensive Training des Klassifikators selbst. Zusammenfassend sollte angemerkt werden, dass die nahezu uner-

[2] Randnotiz: Im Rahmen des „Summer Vision Projects" des Massachusetts Institute of Technology (MIT) im Jahre 1966 veranschlagte S. Papert [Papert, 1966] für eine solche Aufgabe eine Zeitraum von einem Monat für seine Studenten. Die Zielvorgabe für den zweiten Monat war die Erweiterung des entwickelten Verfahrens auf komplexere Umgebungen, Objekte und Oberflächentexturen. Die Tatsache, dass sich im Jahr 2010 ein großer Teil der Vision-Community weltweit mit der gleichen bzw. einer sehr ähnlichen Aufgabenstellung wie die Studenten von S. Papert im Jahr 1966 befasst, zeigt, dass die Segmentierung einer Szene keinesfalls ein gelöstes Problem darstellt.

schöpfliche Menge an Objektklasssen eine generelle Beschreibung von Objekten auf der Basis von objektspezifischen Merkmalen – selbst für Objekte der gleichen Kategorie – sehr schwierig macht [Biederman u. a., 1982]. Als Konsequenz daraus basieren die meisten Ansätze auf sehr restriktiven, einer anwendungsspezifischen Objektklasse zugeschnittenen, Beschreibungen.

Bei den *datengetriebenen Ansätzen* liegt der Schwerpunkt auf einer schnellen Auswertung der Bilddaten selbst. Wünschenswert ist eine möglichst generelle Beschreibung der zu detektierenden Objekte. Die hier zugrunde liegenden Merkmale und Verfahren sind meist sehr einfach und dadurch effizient zu implementieren. Bis auf wenige Ausnahmen werden Merkmale für die Generierung von Objekthypothesen direkt aus dem Bild extrahiert und verarbeitet. Durch den geringen Abstraktionsgrad der Daten muss die Beschreibungskraft des Merkmals bezüglich des Objekts als gering eingestuft werden. Die damit verbundene hohe Fehldetektionsrate ist in den meisten Fällen akzeptabel, da fehlerhafte Objekthypothesen durch die nachgeschaltete höhere Bildinterpretation gelöscht werden können.

Das Ziel dieser Arbeit ist eine möglichst allgemeingültige und minimal einschränkende Beschreibung der Szene. Dabei wird ein einzelnes Objekt als eine räumlich und zeitlich gleichförmig bewegte Gruppierung von Szenenpunkten beschrieben. Bei der Bildung einer geeigneten Repräsentationen der dreidimensionalen Umgebung wird eines der elementarsten Merkmale genutzt, das der Mensch zur Interpretation einer Szene und deren Inhalte benutzt: die *Bewegung*. Aus einer Folge von Bildern wird für jeden Bildpunkt die Bildverschiebung bestimmt, die durch die Bewegung des projizierten Szenepunktes in der Bildebene entsteht. Eine fundamentale Annahme ist dabei die Starrheit der Szene und der darin befindlichen Objekte. Die Welt wird somit als eine Menge sich unabhängig bewegender Objekte aufgefasst, deren unterschiedliche räumliche Bewegung zu einer charakteristischen Grauwertverteilung in den einzelnen Ansichten zeitlich und räumlich versetzter Kamerabilder führt. Das mittlere Bild auf der rechten Seite von Abbildung 1.1 zeigt das endgültige Ergebnis der hier vorgestellten Szenensegmentierung. Detektierte Objekte sind rot unterlegt.

Die Szeneninterpretation schließt mit der höheren Bildinterpretation (IV) ab. Diese fasst Verarbeitungsschritte zusammen, die objekt- und zeitübergreifende Zusammenhänge wie z. B. charakteristische Objektkonfigurationen oder zusammenhängende Bewegungsabläufe auswerten. Aufbauend auf den hier vorgestellten Ergebnissen der bewegungsbasierten Szenensegmentierung wurde im Rahmen dieser Arbeit höher abstrahierte Szeneninformation genutzt, um eine Objektklassifikation auf der Basis relationalen Wissens durchzuführen. Eine detaillierte Beschreibung des Verfahrens und die entsprechenden Ergebnisse können u. a. in den Arbeiten von [Lulcheva u. a., 2008; Bachmann u. Lulcheva, 2009b] nachgelesen werden.

## 1.1 Verfahren zur Objektsegmentierung

Zur Einordnung der vorgestellten Objektsegmentierung, werden im Folgenden die hierfür relevanten Verfahren kurz vorgestellt. Die Übersicht ordnet die einzelnen Ansätze nach unterschiedlichen Kriterien ein, wodurch es teilweise zu Mehrfachnennungen kommen kann. Der Klammerausdruck hinter dem jeweiligen Kriterium bezieht sich dabei auf Tabelle 1.1, welche die vorgestellten Ansätze und deren Eigenschaften nochmals zusammenfasst.

Ein für diese Arbeit wichtiges Kriterium bei der Beurteilung eines Verfahrens ist der jeweilige *Grad der Bildsegmentierung* (I). Hierbei kann unterschieden werden zwischen einer Zerlegung nur markanter Merkmalspunkte (S) (engl. *sparse segmentation*) wie z. B. bei [Talukder u. Matthies, 2004; Dang u. Hoffmann, 2005; Badino u. a., 2006], oder aber einer vollständigen, dichten Segmentierung des Bildes (D) (engl. *dense segmentation*) [Cremers u. Soatto, 2005; Brox u. a., 2006; Wedel u. a., 2007, 2009; Agrawal u. a., 2005]. So wird die vollständige Zerlegung einer Szene in Bereiche ähnlich bewegter Bildpunkte in [Shi u. Malik, 1998] durch die Definition eines Bewegungsvektors für jeden Bildpunkt erreicht, welcher die Wahrscheinlichkeitsverteilung der Bildbewegung ausdrückt. Durch die paarweise Auswertung dieser Bewegungsvektoren wird auf der Basis des *Normalized-Cut*-Kriteriums die Szene in Bereiche ähnlicher Bewegung unterteilt. Während die meisten Segmentierverfahren auf der Auswertung vorab berechneter Daten, wie z. B. optische Flussfelder, Disparitätsfelder, etc. aufbauen, leitet dieser Ansatz die Segmentierung direkt aus den Intensitätswerten der Bildsequenz ab, wodurch ein sehr generell einsetzbares Verfahren entsteht. Ohne die Einführung weiteren Bedingungen erscheint dieser Ansatz jedoch für eine Anwendung auf Bildsequenzen mit komplexeren Inhalten nicht geeignet. In [Schindler, 2005] wird ein Verfahren vorgestellt, welches durch Delaunay-Triangulation aus einem Satz von Merkmalspunkten ein Dreiecksnetz erstellt und dieses unter der Annahme ähnlicher Starrkörperbewegung in unabhängig bewegte Objekte zerlegt. Zur Modellierung der räumlichen Konsistenz ähnlich bewegter Merkmalspunkte wird ein Markov-Zufallsfeld verwendet. Aufgrund des meist hohen Rechenaufwandes der Verfahren und einer begrenzten Modellierbarkeit natürlicher Szenen hat sich jedoch die dichte Segmentierung von Bildsequenzen in der mobilen Umfeldwahrnehmung bislang nicht durchgesetzt.

Eine sehr grundlegendes und die Verarbeitungskette maßgeblich beeinflussendes Kriterium, ist die *Art des Bildaufnahmesystems* (II). Hierbei kann man zwischen Verfahren unterscheiden, die eine Bildsequenz monoskopisch (m) auf der Basis einer einzelnen Kamera auswerten, ober aber die Ansichten mehrerer räumlich starr versetzter Kameras (s) nutzen. Beeindruckende

| Referenz | (I) | | (II) | (III) | | Bemerkung |
|---|---|---|---|---|---|---|
| | D | S | Kamera-system | Eigen-bewegung | Objekt-bewegung | |
| [Feng u. Perona, 1998] | | x | m | - | 3D, starr | ortsfest |
| [Badino u. a., 2006] | | x | s | 3D,starr | d | - |
| [Wedel u. a., 2007] | x | | m | - | d | - |
| [Talukder u. Matthies, 2004] | | x | s | 3D | d | - |
| [Wedel u. a., 2009] | x | | s | 3D | d | - |
| [Csurka u. Bouthemy, 1999] | | x | m | 2D | d | - |
| [Mansouri u. Konrad, 2003] | x | | m | - | 2D,affin | +Kontur |
| [Mueller u. a., 2008] | | x | m | - | - | - |
| [Chang u. a., 1997] | x | | m | - | 2D,affin | - |
| [Franke u. Heinrich, 2002] | | x | s | - | - | - |
| [Agrawal u. a., 2005] | x | | s | 3D, starr | d | - |
| [Cremers u. Soatto, 2005] | x | | m | 2D | d | +Kontur |
| [Brox u. a., 2006] | x | | m | 2D,affin | 2D,affin | +Kontur |
| [Schindler, 2005] | x | x | m | 3D,starr | 3D,starr | Delaunay |
| [Schoenemann u. Cremers, 2006] | x | | m | 2D,affin | 2D,affin | - |
| [Shi u. Malik, 1998] | x | | m | - | - | ortsfest |
| * | x | | s | 3D, starr | 3D, starr | - |

D/S:dichte/merkmalsbasierte Segmentierung; d:(Objekt-)detektion;
s:stereoskopisch; m:monokular; *:hier vorgestellter Ansatz

**Tabelle 1.1**: Übersicht relevanter Objektsegmentierungsverfahren.

Ergebnisse bei der Auswertung monokularer Bildfolgen werden z. B. in [Cremers u. Soatto, 2005; Brox u. a., 2006] vorgestellt. Bei der Bestimmung eines dichten Bewegungsfeldes wird dabei auf die Grundlagen der Variationsrechnung aufgebaut. Teilweise wird hier die Annahme stückweiser Glattheit der Bewegung um ein parametrisches Modell der Objektbewegung erweitert, was zu einer segmentweise parametrischen Darstellung des Bewegungsfeldes führt. Eine Segmentierung der Szene erfolgt dann mit Hilfe der Niveaumengenmethode (engl. *level-set-method*), welche zusätzlich die Modellierung der Objektkontur erlaubt. Mit Hilfe eines kalibrierten Stereokamerasystems wird in [Huguet u. Devernay, 2007; Wedel u. a., 2009] eine dichte Szenensegmentierung aufgrund des sogenannten Szenenflusses durchgeführt. Der Szenenfluss repräsentiert dabei den dreidimensionalen Bewegungsvektor für jeden Bildpunkt im Bildraum, d. h. die Kombination des optischen Flussfeldes und der Disparitätsänderung zwischen zeitlich aufeinanderfolgenden Bildern. Dem Vorteil der subpixelgenauen Bestimmung dichter Bewegungsfelder bei Variationsansätzen steht der hohe Rechenaufwand bei der Bestimmung eines meist komplexen Systems von Differentialgleichungen gegenüber. Eine weitere Einschränkung stellt die begrenzte Anwendbarkeit differentieller Verfahren bei größeren Bildbewegungen dar.

Aufgrund der engen gegenseitigen Verknüpfung der Segmentierung mit der Bewegungsschätzung, erscheint weiterhin eine Einteilung der Verfahren entsprechend dem verwendeten *Bewegungsmodell* (III) sinnvoll. Durch die Einführung von Modellannahmen können ausgedehnte Bildbereiche, also mehrere lokal geschätzte Bildpunktverschiebungen, einem gleichförmig bewegten Objekt zugewiesen werden. Hierbei wird der Tatsache, dass die im Bild beobachtbare Bewegung eine Folge von Bewegungs- und Geometrieeigenschaften widerspiegelt, unterschiedlich stark Rechnung getragen. Bezüglich der Segmentierung einer Szene kann hierbei zwischen einer Bewegungsdetektion (d) oder aber einer Segmentierung auf Basis einer quantitativen Schätzung der Bewegung für jedes Objekt unterschieden werden. So werden z. B. bei [Talukder u. Matthies, 2004; Agrawal u. a., 2005; Wedel u. a., 2007; Klappstein u. a., 2008] die einzelnen Bildpunkte einer Bildsequenz lediglich als eigenbewegt oder fremdbewegt klassifiziert. Ungeachtet der relativen Position in der Szene, werden unabhängig bewegte Objekte dabei zu einem Segment zusammengefasst. Einzelne Objekte können aus diesem Segment mit aufwändigen und meist heuristisch motivierten Verfahren extrahiert werden. Im Gegensatz dazu wird in den Arbeiten von [Chang u. a., 1997; Feng u. Perona, 1998; Schoenemann u. Cremers, 2006] auf der Basis parametrischer Bewegungsmodelle die Bewegung unabhängig bewegter Objekte geschätzt und daraus eine Segmentierung abgeleitet. Ein Vergleich mit dem hier entwickelten Ansatz fällt schwer, da die Leistungsfähigkeit der Verfahren meist nur in einfachen Laborumgebungen und in der Simulation gezeigt wird. Tabelle 1.1 gibt nochmals eine Übersicht der erwähnten Verfahren und ordnet sie hinsichtlich der vorgestellten Kriterien ein.

## 1.2 Eigener Ansatz

Die Gegenüberstellung der oben aufgeführten Verfahren macht deutlich, dass zum gegenwärtigen Zeitpunkt nur in sehr eingeschränktem Umfang bewegungsbasierte Segmentierverfahren existieren, die eine bildpunktgenaue Objektdetektion im Bereich der videobasierten Umfelderfassung durchführen. Die meisten bestehenden Verfahren generieren Objekthypothesen hierbei entweder durch die Auswertung nur einer kleinen Untermenge der Bilddaten auf der Basis bestimmter Deskriptoren oder aber im gesamten Bild auf stark gefilterten und vorverarbeiteten Daten. Die in dieser Arbeit vorgestellte Szenensegmentierung kombiniert beide Ansätze in einem neuen Verfahren und gelangt dadurch zu einer vollständigen Beschreibung der Szene, bei der die einzelnen Modellparameter robust durch merkmalsbasierte Verfahren bestimmt werden. Hieraus ergeben sich folgende Vorteile gegenüber bestehenden Verfahren:

- ◇ Die Qualität der Segmentierung wird durch ein Gütemaß beschrieben, welches an jeder Bildposition definiert ist und somit eine *ganzheitliche Bewertung* der Schätzungen für die gesamte Szene erlauben. Die einzelnen Teilaufgaben der Szenenzerlegung, bestehend aus der dreidimensionalen Rekonstruktion, Bewegungsschätzung und Segmentierung, werden dabei *direkt auf den Grauwerten* der Bildsequenz durchgeführt, anstatt auf vorverarbeiteten Daten. Zusätzlich zum Bewegungsmerkmal wird hierbei die Information einer kalibrierten Stereokamera genutzt.

- ◇ Im Unterschied zu den meisten Verfahren aus der Literatur wird die Bewegung effizient durch ein *parametrisches Bewegungsmodell* beschrieben. Die dreidimensionale Szenenbewegung wird dabei durch eine Menge unabhängig bewegter, starrer Objekte beschrieben. Die Modellparameter werden robust im Rahmen einer *merkmalsbasierten Bewegungschätzung* nachgeführt.

- ◇ Die gegenseitige Abhängigkeit von Szenensegmentierung und Bewegungsschätzung wird durch einen *iterativen Optimierungsprozess* gelöst, wobei das Segmentierungsergebnis in Form eines *probabilistisch motivierten Assoziationsgewichtes* in die Schätzung der Modellparameter jeder einzelnen Objekthypothese eingeht.

- ◇ Inspiriert durch den Wahrnehmungsprozess höher entwickelter Lebewesen werden bei der Modellbildung einige der wichtigsten, dort verwendeten Prinzipien genutzt [Palmer, 1999]: neben der *Ähnlichkeit* der Merkmalsattribute zum Erkennen von Gruppierungen wird weiterhin auch die räumliche und zeitliche *Verbundenheit* zusammengehöriger Szenenbereiche explizit im Modell berücksichtigt.

- ◇ Einen besonderen Charme hat das probabilistische Modell der Szenensegmentierung durch seine *einfache und intuitive Beschreibung*, mit welcher die Integration weiterer Szeneninformationen aus anderen Verarbeitungsebenen oder auch die Kombination mit anderen Verfahren einfach möglich ist. Die hierfür in [Bachmann u. Balthasar, 2008; Bachmann u. Lulcheva, 2009b] durchgeführten Untersuchungen haben das hohe Potenzial einer solchen Beschreibung gezeigt.

Bei der Segmentierung der Bildfolge wird jedem Bildpunkt die Objekthypothese mit den am besten passenden Modellparametern zugewiesen. Das Ergebnis ist eine vollständige Aufteilung des Bildes in sich nicht überlappende Bildbereiche. Jeder dieser Bereiche kann eindeutig als unabhängig bewegtes Objekt interpretiert werden. Durch das parametrische Objektmodell ist es weiterhin möglich, die wesentlichen Szeneninhalte kompakt durch wenige Parameter zu beschreiben und

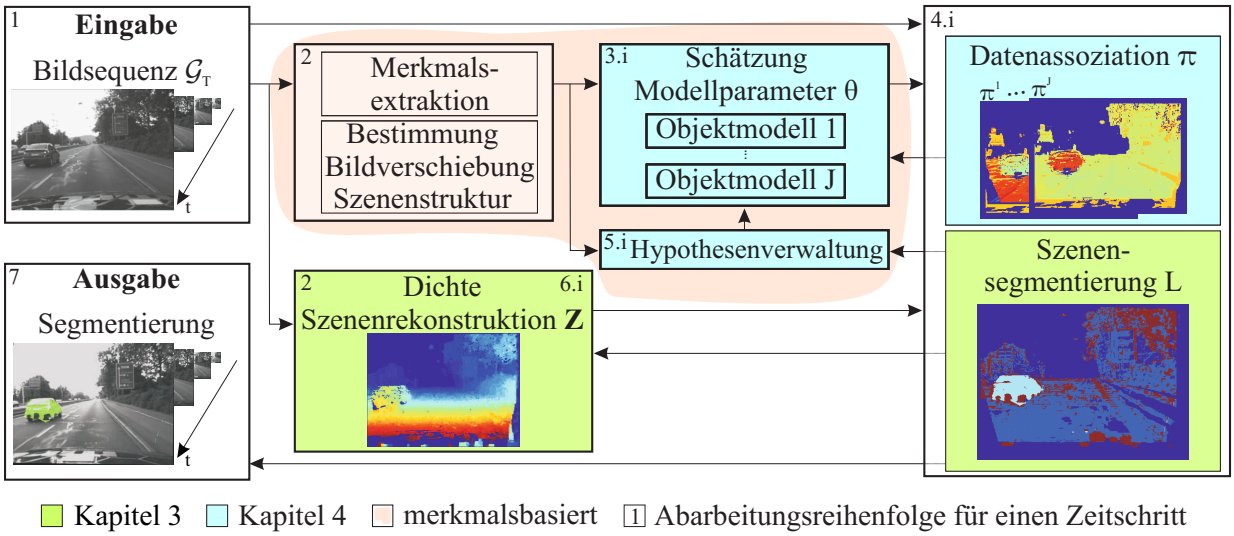



**Abbildung 1.2**: Systemarchitektur der Szenensegmentierung zur bildpunktgenauen Detektion unabhängig bewegter Objekte aus Stereobildfolgen. Das Verfahren besteht aus einer – sich gegenseitig beeinflussenden – Bestimmung der dreidimensionalen Szenenbewegung und Struktur. Daraus abgeleitet ergibt sich die Szenensegmentierung. Die Nummern in den Blöcken markieren, beginnend bei 1, die Abarbeitungsreihenfolge der einzelnen Module zu einem festen Zeitpunkt. Die zusätzlich mit „i" beschrifteten Blöcke werden in alternierender Reihenfolge mehrmals nacheinander pro Zeitschritt ausgeführt.

in der klassischen „*bounding box*"- Darstellung an andere Verfahren weiterzureichen. Neben den oben genannten Teilaufgaben beinhaltet das Verfahren auch eine kontinuierliche Anpassung der optimalen Anzahl von Objekthypothesen auf Basis der aktuellen Szenensegmentierung. Der Einsatz aktueller Optimierungsverfahren macht das Verfahren effizient in Bezug auf Speicherbedarf und Rechenaufwand. Abbildung 1.2 zeigt die einzelnen Komponenten des Systems und wie sie zusammenarbeiten.

## 1.3 Aufbau der Arbeit

In **Kapitel 2** werden die mathematischen und physikalischen Grundlagen der Szenensegmentierung vorgestellt. Das Bildaufnahmesystem wird dabei als ideale Lochkamera modelliert. Neben der Auswertung monokularer Bildfolgen steht in dieser Arbeit zusätzlich die Information eines vollständig kalibrieren Stereokamerasystems zur Verfügung, womit eine nahezu dichte Rekonstruktion der Szene möglich ist. Weiterhin wird eine Übersicht und Einordnung relevanter Mo-

delle und Ansätze zur Bestimmung der dreidimensionalen Bewegung aus Bildfolgen gegeben. Basierend auf der Bewegungsanalyse werden Bildbereiche durch einen Labelprozess zusammengefasst und als gleich bewegte Objekte interpretiert. Zur stochastischen Modellierung räumlicher und zeitlicher Abhängigkeiten werden Markov-Ketten und Markov-Zufallsfelder eingeführt. Die Lösung der Segmentieraufgabe erfolgt durch das globale Optimierungsverfahren der minimalen Graphenschnitte, welches abschließend eingeordnet und näher vorgestellt wird.

**Kapitel 3** befasst sich mit der formalen Beschreibung des verwendeten Objekt- und Szenenmodells. Zur Beschreibung des direkten Fahrzeugumfeldes wird das geometrische Modell einer Ebene eingeführt. Die sich hieraus ergebende Fahrbahnschätzung wird genutzt, um eine Vorsegmentierung der Szene durchzuführen. Für die verbleibenden Bereiche wird die Szenensegmentierung als Schätzproblem im Bayes'schen Sinn formuliert. Die so beschriebene Schätzaufgabe besteht, neben der Bestimmung der Segmentierung und der Szenenstruktur, außerdem noch aus der Schätzung der dreidimensionalen Bewegung, Lage und Ausdehnung eines Objekts. Ein probabilistisches Modell der Szenensegmentierung wird aufgebaut, wobei die starken statistischen Bindungen in räumlicher und zeitlicher Richtung explizit durch die Verwendung eines Markov-Zufallsfeldes berücksichtigt werden. Abschließend werden die einzelnen Teile zu einem Gesamtmodell zusammengefasst.

Da die gleichzeitige Schätzung aller Größen des in Kapitel 3 vorgestellten Modells als nicht praktikabel erscheint, wird in **Kapitel 4** ein Verfahren vorgestellt, welches die einzelnen Komponenten in einem iterativen Prozess unabhängig voneinander optimiert. Für die Bewegungsschätzung wird ein Kalman-Glätter mit einer probabilistisch gewichteten Zuweisung der Beobachtungen vorgestellt. Die sich stetig ändernde Anzahl von Objekten wird durch die Auswertung des aktuellen Segmentierergebnisses kontinuierlich aktualisiert.

Das vorgestellte Verfahren wurde an realen und synthetischen Bildsequenzen erprobt und bewertet. **Kapitel 5** zeigt die Ergebnisse der Szenensegmentierung am Beispiel typischer Verkehrsszenarien. Neben der Auswertung der Module Bewegungschätzung, Szenenrekonstruktion und Bildsegmentierung unabhängig voneinander, wird auch der Einfluss der einzelnen Schätzgrößen auf das Gesamtergebnis untersucht und bewertet. Es wird gezeigt, dass eine probabilistische Datenassoziation der binären Zuweisung von Beobachtungen bei der Bewegungsschätzung deutlich überlegen ist.

**Kapitel 6** fasst die wesentlichen Erkenntnisse und Ergebnisse dieser Arbeit nochmals zusammen. Mit Blick auf die zukünftige Entwicklung im Bereich der bildbasierten Szenenanalyse wird das hohe Potenzial der vorgestellten Szenensegmentierung nochmals aufgezeigt und mögliche nächste Schritte daraus abgeleitet.

KAPITEL 2

# Grundlagen der Szenensegmentierung

In diesem Kapitel werden die für das weitere Verständnis der Szenensegmentierung notwendigen mathematischen und physikalischen Grundlagen vorgestellt. Zur Beschreibung des Bildaufnahmesystems wird in dieser Arbeit das Modell einer idealen Lochkamera benutzt. Neben der monokularen Bildfolge wird zusätzlich die Information eines vollständig kalibrierten Stereokamerasystems ausgewertet, womit eine nahezu vollständige 3D Rekonstruktion der statischen Szene möglich ist. Das für die Bewegungsschätzung genutzte parametrische Modell eines, sich im Raum bewegten, starren Körpers wird vorgestellt und eingeordnet. Aufbauend auf den Ergebnissen der Szenenrekonstruktion und der Bewegungsschätzung wird die Szene schließlich in eine Menge physikalisch relevanter Objekte zerlegt. Hierfür wird ein Labelprozess eingeführt, der ähnlich bewegte Szenenbereiche zusammenfasst. Zur stochastischen Modellierung örtlicher und zeitlicher Abhängigkeiten dieses Segmentierprozesses werden Markov-Zufallsfelder eingeführt. Die Lösung der Segmentieraufgabe erfolgt mit Hilfe von Graphenschnittverfahren, welche abschließend näher vorgestellt werden.

## 2.1 Kameramodellierung

Der Weg vom Objekt in der Szene zum digitalen Bild im Speicher eines Bildaufnahmesystems lässt sich nach [Jähne, 2005] in drei Schritte zerlegen: (i) Sichtbarmachung, (ii) Abbildung und (iii) Digitalisierung. Betrachtet man das Aufnahmesystem als passiven Strahlungsaufnehmer, werden Objekte sichtbar durch Reflexion, Brechung und Streuung des Lichtes der realen Welt. Die physikalischen Objekteigenschaften spielen dabei gleichermaßen eine Rolle wie die Beleuchtungsverhältnisse der Szene. Im Rahmen dieser Arbeit werden Objekte der realen Welt als opak bezüglich der gemessenen Strahlung angenommen. Somit

kann jedem Punkt $\mathbf{x} = (x,y)^{\mathrm{T}}$ in der Bildebene eindeutig ein Punkt $\mathbf{X} = (X,Y,Z)^{\mathrm{T}}$ der realen Welt zugeordnet werden. Die Abbildung der Welt auf der Sensorebene in Form einer Folge von Bildern wird im kontinuierlichen Raum durch $\mathcal{G}_T^{\mathrm{k}} = \{\mathrm{g}^{\mathrm{k}}(\mathcal{X},t|t \in \mathcal{T} = \{0,\ldots,T\})\}$ beschrieben, wobei $t$ die zeitliche Dimension des Signals ausdrückt. Für die auf dem Bildsensor auftreffende Lichtleistung werden gleichermaßen die Begriffe Intensitätswert und Grauwert verwendet.

### 2.1.1 Zentralprojektion

Zur Charakterisierung einer Abbildung der 3D Welt auf einen 2D Aufnehmer wird in diesem Abschnitt das Projektionsmodell einer Kamera vorgestellt. Es besteht aus der sog. perspektivischen Projektion, welche auf dem Modell einer idealisierten Lochkamera basiert. Im Modell wird das Loch als punktförmig angenommen und als Projektionszentrum bezeichnet, da alle in die Kamera einfallenden Sichtstrahlen sich hier schneiden. Ein 3D Szenenpunkt wird durch die Lochblende[1] auf einen einzelnen Bildpunkt der Bildebene abgebildet. Wählt man zur Beschreibung der Weltkoordinaten $\mathbf{X}$ ein Kamerakoordinatensystem mit dem Ursprung im Projektionszentrum $\mathbf{C}$, ergibt sich aus dem Strahlensatz die *Projektionsgleichung*

$$\frac{x'}{\mathrm{f}} = -\frac{X}{Z}, \qquad \frac{y'}{\mathrm{f}} = -\frac{Y}{Z}, \qquad \text{bzw.} \quad \lambda' \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \quad \text{mit} \quad \lambda' \in \mathbb{R}. \tag{2.1}$$

Die $Z$-Achse[2] ist dabei senkrecht zur Bildebene orientiert und die 2D Bildkoordinaten $\mathbf{x}' = (x',y')^{\mathrm{T}}$ sind jeweils antiparallel zur $X$- bzw. $Y$-Achse ausgerichtet. f steht für die Brennweite der Kamera. Diese einfache Gleichung zeigt auf, dass eine Kamera nur Verhältnisse zwischen 3D Koordinaten bestimmen kann, d. h. sie ist maßstabsblind. Absolute Entfernungsangaben lassen sich aus Kamerabildern nur dann bestimmen, wenn zusätzlich eine beliebige Länge in der realen Welt bekannt ist. Explizit soll hier auf die Möglichkeit zur Maßstabsrekonstruktion für statische Szenen mit Hilfe einer kalibrierten Stereoanordnung hingewiesen werden, die im nächsten Abschnitt näher erläutert wird. Anstelle der Bildkoordinaten in der Ebene des Strahlungsaufnehmers wählt man mathematisch eleganter eine dazu parallele Bildebene im Abstand $\mathrm{f} = 1$ vor der Lochblende. Man bezeichnet dies in verallgemeinerten Bildkoordinaten definierte virtuelle Bild, als das Bild einer *kalibrierten*

[1] In der Praxis wird diese Blende durch eine Optik ersetzt, die ein lichtstärkeres Bild erzeugt.

[2] im Weiteren auch als optische Achse bezeichnet

*Kamera* mit der Projektionsgleichung

$$\frac{x'}{\mathrm{f}} \mapsto x \qquad \frac{y'}{\mathrm{f}} \mapsto y\,, \qquad \text{bzw.} \quad \lambda \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \lambda \mathsf{x} = \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \text{ mit } \lambda \in \mathbb{R}\,. \tag{2.2}$$

mit homogenen Koordinaten $\mathsf{x} = (x, y, 1)^{\mathrm{T}}$. Das resultierende geometrische Kameramodell ist in Abbildung 2.1(a) dargestellt.

Im Anschluss an die Projektion werden die normalisierten Bildkoordinaten üblicherweise in Rechnerkoordinaten $\mathsf{x}_R$ überführt, was durch eine affine Transformation der Form

$$\mathsf{x}_R = \mathbf{K}\mathsf{x}\,, \quad \text{mit} \quad \mathbf{K} = \begin{bmatrix} \mathrm{f} & \alpha & x_0 \\ 0 & \beta\mathrm{f} & y_0 \\ 0 & 0 & 0 \end{bmatrix} \tag{2.3}$$

erreicht werden kann. $\mathbf{K}$ wird als *Kalibriermatrix* bezeichnet, welche die intrinsischen Kameraparameter beinhaltet. Neben der Brennweite sind dies der *Bildhauptpunkt* $\mathbf{c} = (x_0, y_0)^{\mathrm{T}}$, die *Scherung* $\alpha$, die den Winkel zwischen $x$- und $y$-Achse der Bildebene quantifiziert und das *Seitenverhältnis* $\beta$ zwischen der vertikalen und horizontalen Seitenlänge eines Bildpunktes. In dieser Arbeit soll vereinfachend von quadratischen Bildpunkten ausgegangen werden, womit $\alpha = 0$ und $\beta = 1$ gesetzt werden kann.

Ist das Weltkoordinatensystem weiterhin nicht an der Kamera orientiert, sondern anwendungsorientiert um die $3 \times 3$-Rotationsmatrix $\mathbf{R}$ gedreht und um den $3 \times 1$-Translationsvektor $\mathbf{T}$ verschoben, ergibt sich für einen beliebigen Weltpunkt $\mathsf{X}_W = (X_W, Y_W, Z_W, 1)^{\mathrm{T}}$ in homogenen Koordinaten folgendes, lineare Gleichungssystem:

$$\mathsf{X} \cong \mathsf{M}\mathsf{X}_W\,, \quad \text{mit} \quad \mathsf{M} = \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ 0 & 1 \end{bmatrix}. \tag{2.4}$$

Die sog. extrinsische Kalibriermatrix $\mathsf{M}$ enthält hierbei die sechs Freiheitsgrade einer starren Bewegung im 3D Raum. Zusammengefasst ergibt sich die Abbildung eines Punktes $\mathsf{X}_W$ in 3D Weltkoordinaten auf 2D Rechnerkoordinaten $\mathsf{x}_R$ als, in homogenen Koordinaten, lineare Abbildung

$$\mathsf{x}_R = \begin{bmatrix} u_R \\ v_R \\ w_R \end{bmatrix} \cong \mathbf{P}\mathsf{X}_W\,, \quad \text{mit} \quad \mathbf{P} = \mathbf{K}\,[\mathbf{R}, \mathbf{T}]\,. \tag{2.5}$$

$\mathbf{P}$ stellt eine $3 \times 4$-Matrix dar, die als *Projektionsmatrix* bezeichnet wird [Hartley u. Zisserman, 2004]. Bei Kenntnis der Parameter in $\mathbf{P}$, liefert (2.5) eine Berechnungsvorschrift, die für den idealen Fall eine Abbildung beliebiger Raumpunkte auf die Bildebene beschreibt. Nachfolgend werden an einigen Stellen die euklidischen Koordinaten eines projizierten Bildpunkts benötigt, wofür die abkürzende Schreibweise

$$\mathbf{x}_R = \Pi(\mathbf{X}_W) = \begin{bmatrix} u_R/w_R \\ v_R/w_R \end{bmatrix} \tag{2.6}$$

für die Berechnung der Bildposition eines Raumpunktes $\mathbf{X}_W$ nach (2.5) definiert wird. Die Inversion von (2.6) wird in dieser Arbeit ebenfalls benötigt und kann bei bekannter Tiefe $Z_W$ durch die Funktion $\Pi^{-1}(\mathbf{x}_R, Z_W)$ ausgedrückt werden. Eine nähere Beschreibung hierzu ist in der hervorragenden Arbeit von [Dang, 2007] zu finden. Die in der Praxis auftretenden Abweichungen vom idealen Lochkameramodell durch Verzeichnungseffekte der Optik werden durch gängige Verfahren und Modelle [Schreer, 2005; Stiller u. a., 2009] korrigiert. Zusätzlich wird die Annahme einer ausreichenden Tiefenschärfe getroffen.

Während die eben beschriebene Projektion eines Szenenpunktes auf die Bildebene ein in Ort, Zeit und Amplitude kontinuierliches Signal erzeugt, wird für die Signalauswertung der Bildfolgen im Rechner ein digitales Signal benötigt, d. h. die Bilder müssen durch *Abtastung* und *Quantisierung* diskretisiert werden. Die Ortsdiskretisierung erfolgt dabei bereits durch das *Sensorraster* auf dem Bildaufnehmer $\mathcal{X} = \{(u,v) \in \mathbb{N}^2 | 0 \leq u < U, 0 \leq v < V\}$ der Dimension $U$ in horizontaler und $V$ in vertikaler Richtung. Da eine solche geordneten Indexierung jedoch die weitere Beschreibung deutlich erschweren würde, wird aus Gründen der einfacheren Notation $\mathcal{R} = \{1, 2, \ldots, N\}$ gewählt, wobei $N = U \cdot V$ die Anzahl der Bildpunkte ist. Aliasingeffekte, die durch Verletzung des Abtasttheorems infolge einer starken Unterabtastung des kontinuierlichen Bildsignals entstehen, werden in dieser Arbeit als gering eingestuft. Für eine regelmäßige Abtastung des Bildsignals im Abstand $\Delta x = \Delta y = \Delta t = 1$ in örtlicher und zeitlicher Richtung ergibt sich schließlich für die quantisierte Bildfolge $\mathcal{G}_T = \mathcal{G}_T^{\mathrm{k}} \cdot \sum_{u,v,w \in \mathbb{Z}} \delta(x-u, y-v, t-w)$. Ein einzelnes Bild der Bildfolge wird im Weiteren auch mit $\mathrm{g}_t(\mathcal{X})$ bzw. $\mathrm{g}_t$ bezeichnet. Die Grauwerte sind mit 8 bit linear quantisiert, womit sich für einen einzelnen Bildpunkt $\mathbf{x}$ ein diskreter Wertevorrat von $\omega_{\mathrm{g}} = \{0, \ldots, 2^8 - 1\}$ ergibt.

### 2.1.2 Modell einer Stereokamera

Bei der Projektion eines 3D Szenenpunktes auf die 2D Bildebene handelt es sich um eine nicht bijektive Abbildung, d. h. es existiert keine eindeutige inverse Abbil-

dung der Bildinformation in die zugrunde liegende Szenengeometrie. Wird dieselbe Szene jedoch aus zwei unterschiedlichen Blickwinkeln betrachtet, kann theoretisch die Tiefeninformation zu allen Bildpunkten bestimmt werden, die in beiden Ansichten existieren. Zur Beschreibung eines solchen Kamerasystems wird das oben eingeführte mathematische Modell einer monokularen Lochkamera um eine zweite Kamera erweitert. Die resultierende Stereoanordnung besteht aus zwei Kameras mit optischen Zentren $\mathbf{C}_\mathrm{r}$ und $\mathbf{C}_\mathrm{l}$, die in ihrem Erfassungsbereich denselben Szenenpunkt $\mathbf{X}$ abbilden. Die Indizes l für *links* und r für *rechts* werden zur Unterscheidung der beiden Kameras verwendet. Aus praktischen Gründen wird im weiteren Verlauf das Weltkoordinatensystem mit dem Kamerakoordinatensystem der rechten Kamera zusammengelegt, d. h. die extrinsischen Parameter der rechten Kamera ergeben sich zu $\mathbf{R}_\mathrm{r} = \mathbf{I}$, $\mathbf{T}_\mathrm{r} = \mathbf{0}$, und die der linken Kamera zu $\mathbf{R}_\mathrm{l} = \mathbf{R}$, $\mathbf{T}_\mathrm{l} = \mathbf{T}$. Die jeweiligen Projektionsmatrizen lauten somit $\mathbf{P}_\mathrm{r} = \mathbf{K}_\mathrm{r}[\mathbf{I},\mathbf{0}]$ , bzw. $\mathbf{P}_\mathrm{l} = \mathbf{K}_\mathrm{l}[\mathbf{R},\mathbf{T}]$. Die Positions- und Orientierungsänderung der linken Kamera relativ zur weltfesten, rechten Kamera kann somit durch eine starre Transformation ausgedrückt werden. Die Transformation beschreibt eine Überführung des Szenenpunktes $\mathsf{X}_\mathrm{r} = \mathsf{X}_W = (X_W, Y_W, Z_W, 1)^\mathrm{T}$ hinsichtlich des Kamerakoordinatensystems des rechten Kamerakoordinatensystems in das linke Kamerakoordinatensystem $\mathsf{X}_\mathrm{l} \cong \mathsf{M}_\mathrm{l}\mathsf{X}_\mathrm{r}$. Die Abbildung des Szenenpunktes auf die 2D Rechnerkoordinaten der linken und rechten Kamera kann mit Hilfe von (2.5) beschrieben werden. Für den 3D Verschiebungsvektor $\mathbf{T}$ der Kameras gilt $\mathbf{T} = \mathbf{C}_\mathrm{l} - \mathbf{C}_\mathrm{r}$. Er wird oft auch als Basis mit der Basisbreite $\mathrm{b} = |\mathbf{T}|$ bezeichnet und drückt den Abstand zwischen den optischen Zentren der beiden Kameras aus.

Um mit Hilfe der so beschriebenen Kameraanordnung eine Szene zu rekonstruieren, müssen im einem ersten Schritt die korrespondierenden, d. h. denselben Szenenpunkt abbildenden, Bildpunkte bestimmt werden, um damit in einem zweiten Schritt die Szenentiefe zu rekonstruieren. Sind nun $\mathbf{x}_\mathrm{r}$ und $\mathbf{x}_\mathrm{l}$ Abbildungen desselben Raumpunktes $\mathbf{X}_W$, kann mit Hilfe der sog. *Epipolarbedingung* der Suchaufwand für Stereokorrespondenzen um eine Dimension entlang der Epipolarlinie reduziert werden. Abbildung 2.1(b) verdeutlicht dies grafisch. Mathematisch lautet die Epipolarbedingung $\mathsf{x}_\mathrm{r}^\mathrm{T}\mathbf{E}\mathsf{x}_\mathrm{l}$ , mit $\mathbf{E} = [\mathbf{T}]_\times\mathbf{R}$. Sie beschreibt die geometrische Beziehung zweier korrespondierender Punkte in den beiden Ansichten des Stereokamerasystems in Bildkoordinaten. $\mathbf{E}$ wird als Essentielle Matrix bezeichnet und ist bei kalibrierten Kameras vollständig durch die Position und Orientierung der beiden Kameras zueinander bestimmt [Fischler u. Firschein, 1987]. In der vorliegenden Arbeit wird $\mathbf{E}$ als vollständig bekannt vorausgesetzt, d. h. neben den intrinsischen Kameraparametern ist auch die relative Lage der Stereokameras zueinander gegeben. Die im Allgemeinen schräg verlaufenden korrespondierenden Epipolarlinien in der Bildebene werden durch eine Transformation, *Rektifikation* genannt, virtuell komplanar zueinander ausgerichtet, womit sich eine für die Korres-

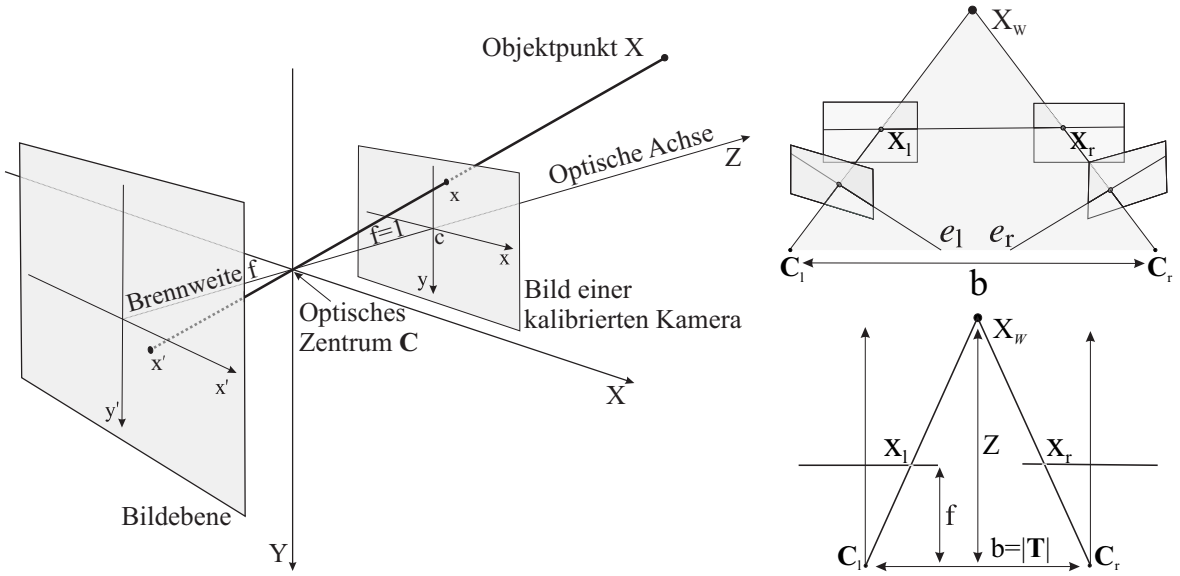



**Abbildung 2.1**: (a) Lochkameramodell der projektiven Abbildung. Ein 3D Objektpunkt $\mathbf{X}$ wird durch die Lochblende auf einen einzigen Bildpunkt $\mathbf{x}'$ bzw. $\mathbf{x}$ der (virtuellen) Bildebene abgebildet. (b) [Angelehnt an [Brown u. a., 2003]] Epipolargeometrie des in Abschnitt 2.1.2 eingeführten Modells einer Stereokamera. Der Suchaufwand für Stereokorrespondenzen wird von der gesamten Bildebene auf eine (Halb-)Gerade reduziert, da $\mathbf{x}_\mathrm{l}$ auf der Halbgeraden liegt, welche durch die von $\mathbf{x}_\mathrm{r}$ bestimmten Epipolarlinie durch den Epipol $e_\mathrm{r}$ gegeben ist.

pondenzsuche deutlich günstigere Struktur ergibt. Eine detaillierte Beschreibung gängiger Rektifizierungsverfahren findet sich in [Trucco u. Verri, 1998]. Nach der Rektifikation erhält man Bilder, die eine Stereoanordnung mit $\mathbf{R}_\mathrm{r} = \mathbf{I}$, $\mathbf{T}_\mathrm{r} = \mathbf{0}$, $\mathbf{R}_\mathrm{l} = \mathbf{I}$, $\mathbf{T}_\mathrm{l} = (\mathrm{b}, 0, 0)^\mathrm{T}$ aufgenommen hätte. Man spricht dann von einer *achsparallelen Stereogeometrie* bzw. von einem rektifizierten Stereosystem. Da Korrespondenzen bei einem achsparallelen Stereosystem in derselben Bildzeile liegen, führt die unterschiedliche Perspektive der Kameras hinsichtlich des Szenenpunktes zu einer rein horizontalen Verschiebung in der Abbildung[3].

[3]Dies kann durch folgende Umformung von (2.1) für den Raumpunkt $\mathbf{X}$ in den jeweiligen Bildkoordinaten unmittelbar erkannt werden

$$\begin{pmatrix} x \\ y \end{pmatrix}_\mathrm{l} = \frac{\mathrm{f}}{Z_\mathrm{l}} \begin{pmatrix} X_\mathrm{l} \\ Y_\mathrm{l} \end{pmatrix} = \frac{\mathrm{f}}{Z_\mathrm{r}} \begin{pmatrix} X_\mathrm{r} + \mathrm{b} \\ Y_\mathrm{r} \end{pmatrix} \qquad \text{und} \qquad \begin{pmatrix} x \\ y \end{pmatrix}_\mathrm{r} = \frac{\mathrm{f}}{Z_\mathrm{r}} \begin{pmatrix} X_\mathrm{r} \\ Y_\mathrm{r} \end{pmatrix}. \tag{2.7}$$

## 2.2 Bewegungsschätzung und Szenenrekonstruktion

Die Auswertung eines kalibrierten Stereobildpaares erlaubt die nahezu vollständige Rekonstruktion der abgebildeten Szene. Da das Bildpaar zeitgleich aufgenommen wird, stellt die Rekonstruktion eine statische Momentaufnahme der Szene dar. Für eine zusätzliche Auswertung der Szenendynamik ist es deshalb notwendig, die Bilddaten auch in zeitlicher Richtung auszuwerten.

### 2.2.1 Schätzung der 3D Bewegung

Die Bewegungsanalyse wertet Bildverbände in zeitlicher Richtung hinsichtlich von Bildmustern aus, die durch Bewegungen in der 3D Szene verursacht werden. Um für diese Größe ein besseres Verständnis zu bekommen, betrachte man die relative Bewegung der Punkte $\{\mathbf{X}_{n,t}|n \in \mathcal{R}\}$ zwischen Kamera und Szene, welche durch ein 3D Geschwindigkeitsvektorfeld[4] $\{\dot{\mathbf{X}}_{n,t}|n \in \mathcal{R}\}$ beschrieben wird. Über die Zeit betrachtet, beschreibt dieses Bewegungsfeld für jeden Bildpunkt eine kontinuierliche Bewegungstrajektorie im Raum. Das projizierte kontinuierliche 2D Bewegungsfeld aller sichtbaren Punkte auf der Sensorfläche ergibt sich dann zu $\{\dot{\mathbf{X}}_{n,t}|n \in \mathcal{R}\} \in \mathbb{R}^3 \mapsto \{\dot{\mathbf{x}}_{n,t}|n \in \mathcal{R}\} \in \mathbb{R}^2$. In zeitdiskreten Bildfolgen wird für die zeitlich korrespondierenden Bildpunkte $\mathbf{x}_t$ und $\widehat{\boldsymbol{\chi}}_{t+\Delta t} = \widehat{\boldsymbol{\chi}}_t$ diese Bewegungsinformation durch den Verschiebungsvektor (engl. *displacement vector*) $\mathbf{d}_t \in \mathcal{D}_t = \{\mathbf{d}_{n,t}|n \in \mathcal{R}\}$ ausgedrückt. Das kontinuierliche Bewegungsfeld $\{\dot{\mathbf{x}}_{n,t}|n \in \mathcal{R}\}$ und das diskrete Verschiebungfeld $\mathcal{D}_t$ unterscheiden sich nur um einen Faktor und können als wechselseitig austauschbare Größen benutzt werden. Im weiteren Verlauf der Arbeit werden daher die Begriffe Bewegungsfeld und Verschiebungfeld gleichermaßen verwendet.

Ziel dieser Arbeit ist es, aus den messbaren Grauwertmustern der Bildfolge, die oben beschriebene Bildbewegung zu bestimmen und daraus eine Segmentierung unabhängig im Raum bewegter Szenenobjekte abzuleiten. Prinzipiell muss dabei bei der Schätzung eines ortsdiskreten dichten Bewegungsfeldes ein Zustandsraum mit insgesamt $\mathbb{R}^{|\mathcal{R}|}$ möglichen Bewegungsvektoren untersucht werden. Dieser Zustandsraum kann jedoch stark eingeschränkt werden, da nicht alle Bewegungsfelder aus diesem Raum in der realen Welt ähnlich häufig auftreten. So können Annahmen für das zu bestimmende Bewegungsfeld getroffen werden, die zu erwartende Geometrie- und Bewegungseigenschaften der realen Welt abbilden und somit den Schätzprozess stützen. Bei der Bewegungsschätzung werden diese Annahmen in Form von Bewegungsmodellen beschrieben, welche nach

[4] Im weiteren Verlauf dieser Arbeit wird $\{\dot{\mathbf{X}}_{n,t}|n \in \mathcal{R}\}$ auch als Bewegungsfeld bezeichnet.

[Bergen u. a., 1992] in drei Gruppen eingeteilt werden können: nicht-parametrisch, quasi-parametrisch und vollständig parametrisch.

*Nicht-parametrische Modelle* umfassen den Bereich der Verfahren, die Bedingungen explizit an das zu schätzende Verschiebungsfeld direkt stellen. Die Verfahren zur Bestimmung der 2D Bildbewegung gründen üblicherweise auf diesen Modellen.

Bei den *quasi-parametrischen Modellen* wird die Bildpunktbewegung als eine Kombination aus einer parametrischen Komponente, welche global auf den gesamten Bildbereich anwendbar ist, und einer nicht parametrischen Komponente, welche sich lokal zwischen den Bildpunkten ändert, beschrieben. Da für die reale Welt eine vollständige Beschreibung der Szene durch einen Satz von Parametern nur eingeschränkt möglich ist, können quasi-parametrische Modelle genutzt werden, um ein parametrisches Bewegungsmodell mit einem lokalen Modell der Szenenstruktur zu kombinieren. Häufig wird hier das Modell der *Starrkörperbewegung* genutzt. Dabei beschreiben die Bewegungsparameter die zeitliche Verschiebung eines Bildpunktes entlang einer Geraden, während der lokal am Bildpunkt vorliegende Wert der Szenentiefe den Verschiebungsvektor eindeutig bestimmt. Die Relativbewegung eines Punktes, der auf einem als starr anzunehmenden Objekt liegt, wird dabei beschrieben durch[5]

$$\dot{\mathbf{X}}_t = \left(\dot{X}, \dot{Y}, \dot{Z}\right)^{\mathrm{T}} = -\mathbf{t}_t - \boldsymbol{\omega}_t \times \mathbf{X}_t \,. \tag{2.8}$$

Relativ zum Koordinatenursprung des Referenzsystems drückt $\boldsymbol{\omega} = (\omega_x, \omega_y, \omega_z)^{\mathrm{T}}$ hier die rotatorische und $\mathbf{t} = (t_x, t_y, t_z)^{\mathrm{T}}$ die translatorische Bewegungskomponente des Starrkörpermodells aus, wie in Abbildung 2.2 skizziert.

Die Bildbewegung $\dot{\mathbf{x}}_t$ als Folge dieser Relativbewegung kann entsprechend den Ausführungen von Longuett-Higgins und Prazdny [Longuet-Higgins u. Prazdny, 1980] durch das Modell instantaner Bewegung[6] mit Hilfe von (2.3) folgendermaßen beschrieben werden

$$\begin{aligned} \dot{\mathbf{x}}_t &= \mathbf{C}_{\boldsymbol{\omega}}\,\boldsymbol{\omega}_t + \frac{1}{Z_t}\mathbf{C}_{\mathbf{t}}\mathbf{t}_t = [\underbrace{\mathbf{C}_{\boldsymbol{\omega}}, \mathbf{C}_{\mathbf{t}}}_{:=\mathbf{C}}](\underbrace{\omega_x, \omega_y, \omega_z, t_x, t_y, t_z}_{:=\mathbf{v}})^{\mathrm{T}}, \text{ mit} \\ \mathbf{C}_{\boldsymbol{\omega}} &= \begin{bmatrix} xy & -\left(1+x^2\right) & y \\ \left(1+y^2\right) & -xy & -x \end{bmatrix}, \; \mathbf{C}_{\mathbf{t}} = \frac{1}{Z_t}\begin{bmatrix} -1 & 0 & x \\ 0 & -1 & y \end{bmatrix}. \end{aligned} \tag{2.9}$$

[5] Ist Zeitindex $t$ für das Verständnis einer bestimmten Größe bzw. Formulierung nicht von Belang, wird im Weiteren zugunsten der besseren Lesbarkeit darauf verzichtet.

[6] Aus dem Englischen *instantaneous motion model*. In der Literatur findet sich hierfür auch der Begriff *differential epipolar constraint*.

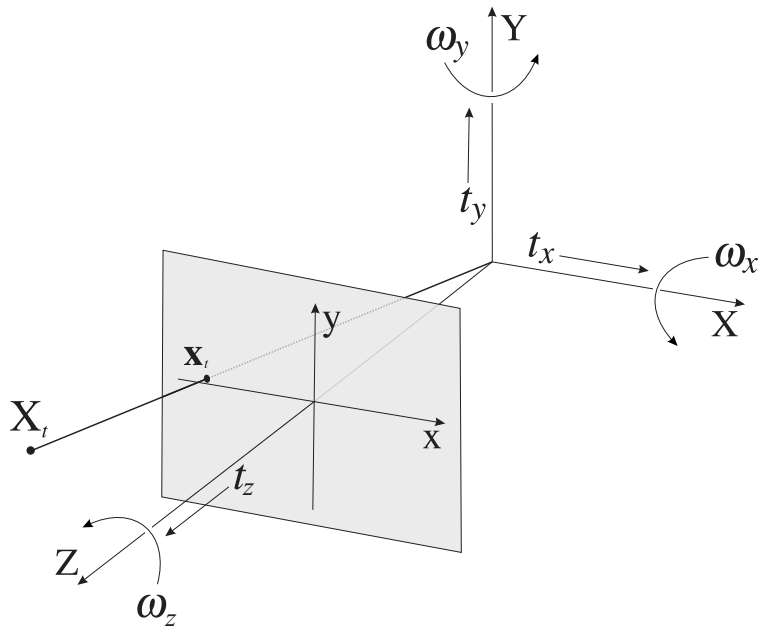



**Abbildung 2.2**: [Angelehnt an [Longuet-Higgins u. Prazdny, 1980]] Die Bildposition $\mathbf{x}_t$ als Zentralprojektion des Raumpunktes $\mathbf{X}_t$ auf die Bildebene. $\boldsymbol{\omega}_t$ und $\mathbf{t}_t$ drücken die relative Rotation und Translation eines Objekts in der Szene aus.

Analog zur Berechnung der instantanen Bildbewegung lässt sich auch ein Modell für die Bildverschiebung im zeitdiskreten Fall aufstellen. So lässt sich die Bewegung zweier korrespondierender Szenenpunkte $\mathbf{X}_t$ und $\widehat{\boldsymbol{\chi}}_{\mathrm{x},t}$ durch

$$\widehat{\boldsymbol{\chi}}_{\mathrm{x},t} = \mathrm{R}_t \mathbf{X}_t + \mathbf{S}_t\,, \quad \text{mit} \quad \mathrm{R}_t = \begin{pmatrix} 1 & -\omega_z & \omega_y \\ \omega_z & 1 & -\omega_x \\ -\omega_y & \omega_x & 1 \end{pmatrix} \tag{2.10}$$

ausdrücken. Hierbei wird die Annahme getroffen, dass zwischen den Aufnahmezeitpunkten der Bilder nur kleine Rotationswinkel auftreten, womit die Rotationsmatrix $\mathrm{R}_t$ durch die Rotationsparameter $\boldsymbol{\omega}_t$ approximiert werden kann [Fang u. Huang, 1984]. $\mathbf{S}_t = (s_x, s_y, s_z)^{\mathrm{T}}$ beschreibt die translatorische Verschiebung des Szenenpunktes im Raum. Mit (2.3) können hieraus die Komponenten des korrespondierenden Bildpunkts $\widehat{\boldsymbol{\chi}}_t$ folgendermaßen berechnet werden:

$$\widehat{\chi}_x = \frac{x - \omega_y + y\omega_z + s_x/Z}{-\omega_y x + \omega_x y + 1 + s_z/Z}\,, \quad \widehat{\chi}_y = \frac{y - \omega_x + x\omega_z + s_y/Z}{-\omega_y x + \omega_x y + 1 + s_z/Z}\,. \tag{2.11}$$

Die Komponenten des Verschiebungsvektors $\mathbf{d}_t = (\widehat{\chi}_x - x, \widehat{\chi}_y - y)^{\mathrm{T}}$ ergeben sich dann zu

$$
\begin{aligned}
d_x &= \frac{xy\omega_x - (1+x^2)\omega_y + y\omega_z + (s_x - s_z x)/Z}{-\omega_y x + \omega_x y + 1 + s_z/Z} \\
d_y &= \frac{(1+y^2)\omega_x - xy_n\omega_y - x\omega_z + (s_y - s_z y)/Z}{-\omega_y x + \omega_x y + 1 + s_z/Z} \; .
\end{aligned}
\tag{2.12}
$$

Obige Gleichung zeigt, dass für den Fall kleiner Tiefenänderungen $s_z$ relativ zur absoluten Tiefe $Z$ und kleiner Rotationswinkel, das Bewegungsfeld approximativ durch das Verschiebungsvektorfeld $\mathcal{D}$ ausgedrückt werden kann. Als zu schätzende Größe ergibt sich, neben den Bewegungsparametern selbst, die Szenenstruktur $Z \in \mathbf{Z}$. Diese kann für beliebige Oberflächen als Funktion $Z(x,y,t)$ der Bildkoordinaten ausgedrückt werden. Ganz allgemein drückt diese Funktion die Oberfläche der realen Szene aus.

Wird zusätzlich zur Bewegung auch die Oberfläche durch einen Satz von Parametern ausgedrückt, ergibt sich ein *vollständig parametrisches Modell*. Diese Modelle beschreiben die Bewegung eines Bildpunktes in einem entsprechend gültigen Bildbereich näherungsweise durch wenige Parameter. Mit einem solches Modell kann auch in schwach bzw. nicht texturierten Bildbereichen eine gültige Bewegungsschätzung erfolgen, vorausgesetzt die Region enthält zumindest einige signifikante Texturmuster [Adiv, 1985].

Die Entscheidung für ein spezielles Modell variiert je nach Anwendungsfall und Szene in Bezug auf Komplexität und Beschreibungskraft. So erscheint es naheliegend, dass mit steigender Anzahl der Parameter bei einem parametrischen Modell auch zunehmend komplexere Bewegungen abgebildet werden können. Als Extremfall kann hier die Parametrierung jedes einzelnen Bildpunktes angesehen werden, was letztendlich zu den nicht-parametrischen Modellen führt. Da in dieser Arbeit Objekte durch ihre ähnliche Bewegung im Raum charakterisiert sind, bieten sich jedoch Modelle mit einer geringen Anzahl von Parametern an. Neben einer effizienten Schätzung der Bewegungskomponenten für einzelne Objekte können durch ein solches Modell außerdem schon während des Schätzprozesses Anforderungen an das Bewegungsfeld gestellt werden, die eine anschließende Interpretation der bewegten Szene deutlich erleichtern.

Im Folgenden werden die Verfahren der modellbasierten Schätzung der 3D Bewegung in zwei Hauptgruppen eingeteilt. Die Taxonomie lehnt sich dabei an die Definition von [Torr u. Zisserman, 2000; Irani u. Anandan, 2000] an: *direkte Verfahren* minimieren hier eine photometrische Größe, die in einer – meist lokalen – Umgebung um den betrachteten Punkt gemessen wird. *Indirekte Verfahren* hingegen minimieren bzw. maximieren ein Maß, welches durch eine Menge von vorher aus dem Bild extrahierten, korrespondierenden Merkmalen selbst definiert wird.

**Direkte Verfahren**

Verfahren aus dieser Gruppe beruhen auf der Auswertung des Grauwertsignals selbst und ggf. seinen Ableitungen. Hierbei kann weiterhin unterteilt werden in die Bereiche der flussbasierten Verfahren, die zur Bestimmung der Bewegungsparameter auf das vorher aus dem Bild extrahierten Verschiebungsfeld aufbauen und Verfahren, die den Parametersatz direkt aus den Grauwerten ableiten. Kann angenommen werden, dass sich die Intensität eines Objektpunktes mit der Zeit nicht ändert, ist auch der Grauwert des auf die Bildebene projizierten Raumpunktes entlang der zugehörigen Bewegungstrajektorie konstant. Für den Fall eines idealen Lambert'schen Strahlers, kann bei konstanter Beleuchtung und ausschließlich translatorischer Bewegung in der Szene die Intensitätsänderung im Bild vollständig durch die im Zeitintervall $\Delta t$ stattfindenden Bewegung $\mathbf{u}_t = \int_t^{t+\Delta t} \dot{\mathbf{x}}_\varsigma d\varsigma$ beschrieben werden. Hieraus ergibt sich für die Grauwerte zeitlich nacheinander aufgenommener Bilder folgende einfache Bedingung:

$$g(\mathbf{x},t) = g(\mathbf{x} + \mathbf{u} \cdot \Delta t, t + \Delta t)\,. \tag{2.13}$$

Für den Fall geringer Intensitätsänderungen in räumlicher und zeitlicher Richtung kann die rechte Seite von (2.13) durch Taylor-Entwicklung erster Ordnung erweitert werden. Dies führt zur bekannten *Kontinuitätsgleichung des optischen Flusses* (engl. *optical flow constraint equation*), die von [Horn u. Schunck, 1981] erstmals eingeführt wurde und aussagt, dass der Bildgradient entlang der Bewegungstrajektorie verschwindet

$$\nabla g_t^T \mathbf{u}_t + \frac{\partial g(\mathbf{x},t)}{\partial t} = 0\,, \text{mit}\;\; \nabla g_t^T = \left( \frac{\partial g(\mathbf{x},t)}{\partial x}, \frac{\partial g(\mathbf{x},t)}{\partial y} \right). \tag{2.14}$$

Die *flussbasierten Verfahren* zur Schätzung der 3D Bewegung bauen indirekt auf diese Gleichung auf. In einem ersten Schritt wird die 2D Bildbewegung bestimmt. Darauf aufbauend wird dann im zweiten Schritt dieses Bewegungsfeld „interpretiert", d. h. die eigentliche Schätzung der Bewegungsparameter durchgeführt. Für ein gegebenes parametrisches Bewegungsmodell bildet jeder Verschiebungsvektor dann eine Bedingung für den Parametervektor. Im Folgenden wird, nach einer kurzen Vorstellung gebräuchlicher Verfahren zur Schätzung der 2D Bildbewegung, eine Auswahl flussbasierter Verfahren zur Schätzung der 3D Bewegung vorgestellt und in die vorliegende Arbeit eingeordnet.

Bestimmung der 2D Bildbewegung

Nach [Barron u. a., 1992] kann die Bestimmung des 2D Bewegungsfeldes *gradienten-* oder *korrespondenzbasiert* durchgeführt werden[7]. Die zu bestimmende Größe $\mathbf{u}_t \in \mathcal{U}_t = \{\mathbf{u}_{n,t} | n \in \mathcal{R}\}$ wird in diesem Zusammenhang als optischer Fluss[8] bezeichnet. Wie bereits erwähnt, ist der optische Fluss grundsätzlich eine nicht direkt messbare Größe. Wenn man also vom optischen Fluss spricht, meint man immer Näherungen. Aus (2.14) ist ersichtlich, dass die Bestimmung des optischen Flusses unterbestimmt ist, da nur eine skalare Bedingung für die gesuchte Größe aufgestellt wird. Zur Berechnung des Flussfeldes $\mathcal{U}_t$ werden deshalb zusätzliche Bedingungen benötigt. Diese werden meist in Form von *Glattheits-* oder *Ähnlichkeitsbedingungen* der zu schätzenden Größe in einer Region um den betrachteten Bildpunkt gestellt. Der Großteil der in den letzten Jahrzehnten entwickelten Verfahren lässt sich dabei in zwei Gruppen unterteilen: Ansätze, die basierend auf der Arbeit von [Horn u. Schunck, 1981] eine *globale* Glattheitsanforderung des zu bestimmenden Verschiebungsfeldes im ganzen Bild definiert oder aber Verfahren, die entsprechend den Ausführungen von [Lucas u. Kanade, 1981] die Ähnlichkeitsbedingungen explizit als konstanten oder stetigen Fluss innerhalb eines Bildbereichs rund um den *lokal* betrachteten Bildpunkt formulieren. Die zur ersten Gruppe gehörenden Variationsverfahren liefern hier gegenwärtig die besten Ergebnisse bezüglich Genauigkeit und Vollständigkeit wie in [Bruhn u. a., 2005] gezeigt wird. In [Huguet u. Devernay, 2007] wird ein Verfahren vorgestellt, welches den klassischen Variationsansatz zur Schätzung des optischen Flusses um eine Dimension im Bildraum erweitert und zusätzlich die Tiefenänderung in der Szene mitschätzt. Hierfür steht ein kalibriertes Stereokamerasystem zur Verfügung. Das Ergebnis dieser Schätzung, ein 3D Verschiebungsfeld für jeden Punkt in der Szene, wird von den Autoren als Szenenfluss (engl. *scene flow*) bezeichnet. In [Wedel u. a., 2008] wird ein ähnlicher Ansatz verfolgt, wobei eine deutliche Komplexitätsreduktion des Problems durch die Trennung der Tiefenschätzung von der Schätzung des Verschiebungsfeldes erreicht wird. Je nach Grad der gestellten Forderungen besteht der Hauptvorteil gradientenbasierter Verfahren in der generellen Anwendbarkeit, da nur minimale Bedingungen an das zu schätzende Bewegungsfeld gestellt werden. Dem steht der Nachteil gegenüber, dass obige Gleichung nur für kleine Verschiebungen gilt. Allgemein werden solche Verfahren daher meist

[7] Die dritte Gruppe der frequenzbasierten Verfahren zur Bestimmung des 2D Bewegungsfeldes ist für diese Arbeit nicht von Belang und wird deshalb hier nicht weiter betrachtet.

[8] In der ursprünglichen, biologisch motivierten Definition des optischen Flusses [Nakayama u. Loomis, 1974] erfolgt die Projektion auf eine kugelförmige Sensorfläche. Im Bereich der maschinellen Bildverarbeitung wird jedoch die zentralperspektivische Projektion auf eine planare Fläche als adäquate Modellierung der Abbildungseigenschaften gebräuchlicher Kameras verwendet.

bei der Bewegungsbestimmung in Bildfolgen eingesetzt, in denen eine geringe Verschiebung der Bildpunkte angenommen werden kann.

Alternativ kann das 2D Verschiebungsfeld auch durch *Korrespondenzverfahren* bestimmt werden. Hierbei kann weiterhin zwischen *merkmalsbasierten* und *regionenbasierten* Verfahren unterschieden werden [Haußecker u. Spies, 1999]. Diese Verfahren sind zur Erfassung größerer Bewegungen meist besser geeignet als gradientenbasierte Verfahren, da die Verschiebung $\mathbf{d}_t = \mathbf{u}_t \cdot \Delta t$ ganzer Bildelemente zweier im Abstand $\Delta t$ aufgenommener Bilder bestimmt wird. Für $\Delta t \rightarrow 0$ geht diese Verschiebung in den optischen Fluss über. Die Bestimmung der Verschiebung erfolgt durch paarweise Zuordnung ähnlicher Bildelemente in Folgebildern. Zur Bewertung der Ähnlichkeit werden dabei meist die Intensitätswerte der einzelnen Ansichten direkt verwendet. Die bedeutendste Gruppe innerhalb der regionenbasierten Zuordnungsverfahren[9] bilden die sog. *Blockzuordnungsverfahren* (engl. *block matching algorithm*). Die Bezeichnung rührt von der Tatsache her, dass ein rechteckiger Bildausschnitt $\mathcal{B}_1 \in \mathcal{X}$ des Referenzbildes $g^1(\mathcal{X},t)$, mit einem Bildausschnitt $\mathcal{B}_2 \in \mathcal{X}$ in Suchbild $g^2(\mathcal{X},t)$ verglichen wird. Gängige Metriken zur Bewertung der Ähnlichkeit an jeder Stelle im Bild sind u. a. die aufsummierte Differenz der Grauwerte $\sum_{\mathbf{x}\in\mathcal{B}_1,(\mathbf{x}+\mathbf{d}_t)\in\mathcal{B}_2} \|g^1(\mathbf{x}) - g^2(\mathbf{x}+\mathbf{d}_t)\|$ innerhalb des Bildausschnitts entsprechend einer Norm oder auch dessen (mittelwertbereinigter) normalisierter Kreuzkorrelationskoeffizient. Im Gegensatz zu den gradientenbasierten Verfahren, werden Korrespondenzen nur innerhalb des diskreten Bildgitters gesucht. Mögliche Positionen können dabei z. B. durch alle ganzzahligen Pixelkoordinaten gegeben sein. Die regionenbasierten Verfahren gehen in die merkmalsbasierten Verfahren über, indem die Korrespondenzsuche auf markante Bildstrukturen wie z. B. Ecken, Linien, etc. im Bild beschränkt wird. Hierauf wird im nächsten Abschnitt näher eingegangen. Die Verwendung markanter Bildstrukturen hat zur Folge, dass homogene Bildregionen, bei denen das Korrespondenzproblem nicht eindeutig gelöst werden kann, prinzipbedingt ausgeschlossen werden. Ein weiterer großer Vorteil merkmalsbasierter Korrespondenzverfahren ist die Möglichkeit der zeitlichen Verfolgung einzelner Merkmalspunkte über längere Bildsequenzen hinweg, womit ein zeitlich stabiles Flussfeld bestimmt werden kann. Zur zeitlichen Verfolgung von 2D Punktmerkmalen wurden in dieser Arbeit der KLT-Tracker (Kanade-Lucas-Tomasi, [Shi u. Tomasi, 1994]) und ein korrelationsbasiertes Verfahren auf der Basis des Harris- [Harris u. Stephens, 1988] und des FAST- (engl. *FAST=Features from Accelerated Segment Test*) [Rosten u. Drummond, 2006] Eckendetektors verwendet.

[9] In der Literatur ist hierfür auch der Begriff der *pixelbasierten Korrespondenzanalyse* gebräuchlich.

Bestimmung der Bewegungsparameter

Mit dem Ziel der Segmentierung einer Szene in unabhängig bewegte Objekte der realen Welt kann die Bestimmung eines, ggf. dichten, optischen Flussfeldes nur als Zwischenschritt aufgefasst werden. Was bleibt, ist das Herauslösen der jeweiligen 3D Objektbewegung aus diesem 2D Vektorfeld. Hierfür kann das Vektorfeld, entsprechend dem gewählten Bewegungsmodell, parametriert und die enthaltenen Bewegungsparameter durch Minimierung des Gütekriteriums

$$\sum_{n \in \mathcal{R}} |\mathbf{u}_n - \hat{\mathbf{u}}_n(\boldsymbol{\omega}, \mathbf{t}, Z_n)| \tag{2.15}$$

bestimmt werden. Ein vielfach verwendetes Bewegungsmodell ist in diesem Zusammenhang das in (2.9) beschriebene Modell der Starrkörperbewegung. Die unbekannte Tiefe $Z$ kann dabei durch algebraische Umformung [Zhuang u. a., 1988; Heeger u. Jepson, 1992; MacLean, 1999; Baumela u. a., 2000] bzw. Rücksubstitution [Bruss u. Horn, 1981] eliminiert werden. In den Arbeiten von [Adiv, 1985; Wang u. Adelson, 1994] wird zur Beschreibung der Szenenstruktur ein einfaches Ebenenmodell verwendet, dessen Parameter zusätzlich mitgeschätzt werden. [Ohta, 2003] beschreibt die Aufgabe der Parameterbestimmung als Maximum-Likelihood-(ML-)Schätzer, wobei die unbekannte Tiefe als stochastische Störgröße modelliert wird. [Mandelbaum u. a., 1999] schlägt ein korrelationsbasiertes Verfahren vor, welches zwar auf (2.9) aufbaut, jedoch auf eine explizite Berechnung des optischen Flussfeldes verzichtet. Stattdessen wird die Bewegungs- und Strukturschätzung direkt auf den Grauwerten des Bildes durchgeführt. Dies führt zu Verfahren, die Bewegung und Struktur schätzen, ohne vorher den optischen Fluss zu bestimmen. In [Irani u. Anandan, 2000] werden zur Lösung dieser Aufgabe, welches in der Literatur allgemein unter dem Begriff „structure-from-motion" (SFM) bekannt ist, verschiedene 2D und 3D Bewegungsmodelle miteinander verglichen und anhand realer Bildsequenzen getestet. [Hanna, 1991] stellt in diesem Zusammenhang ein Verfahren vor, welches das Problem der gleichzeitigen Bestimmung von Bewegung und Struktur aus monokularen Bildfolgen durch Parametrisierung des optischen Flusses und anschließender Integration der Kontinuitätsgleichung des optischen Flusses ebenfalls direkt löst.

Die Aufgabe der Bewegungsschätzung vereinfacht sich merklich, falls ein kalibriertes Stereokamerasystem zur Verfügung steht. So integriert [Stein u. Shashua, 2000] die dritte Ansicht einer Stereokamera und schätzt die Bewegung durch Kombination der Kontinuitätsgleichung des optischen Flusses mit dem geometrischen Modell des trilinearen Tensors. Hieraus ergibt sich die *tensor brightness constraint*, welche die Beziehung zwischen räumlich und zeitlich versetzten Bildgradienten eines Bildpunktes in drei Ansichten

beschreibt. Mit dem Bewegungsmodell aus (2.9) ergibt sich hieraus eine parametrische Bedingungsgleichung für die Grauwerte jedes Bildpunktes. Die zentrale Idee der in [Talukder u. Matthies, 2004; Dang u. Hoffmann, 2005; Horn u. a., 2006; Bachmann u. Dang, 2006a,b; Kitt, 2008] beschrieben Verfahren besteht darin, Merkmalspunkte über mehrere Bilder zu verfolgen und die Bewegungsschätzung durch zeitliche Integration kontinuierlich zu verbessern. Die Lösung der Schätzaufgabe erfolgt rekursiv mit Hilfe eines Kalman-Filters. Als Beobachtung steht den Verfahren das Verschiebungsfeld der Merkmalspunkte, sowie die stereoskopisch geschätzte Tiefe zur Verfügung. Für die Auswertung der Beobachtungsgleichung wird die Tiefe hier als konstant betrachtet. Das *6D-Vision*-Verfahren von [Franke u. a., 2005] verfolgt einen ähnlichen Ansatz, unterscheidet sich jedoch von den oben aufgeführten Verfahren durch die Annahme, dass sich die einzelnen Merkmalspunkte als Teile massebehafteter Körper kurzfristig geradlinig im Raum bewegen. Dies hat zur Folge, dass die Starrheitsbedingung der Szenenpunkte fallengelassen werden kann. Der Zustandsvektor besteht aus den Ortskoordinaten $(X,Y,Z)^{\mathrm{T}}$ der Merkmale sowie den dazugehörigen Geschwindigkeitskomponenten $(\dot{X},\dot{Y},\dot{Z})^{\mathrm{T}}$.

### Indirekte Verfahren

Die indirekten Verfahren zur Bewegungsschätzung gründen meist auf der von [Longuet-Higgins, 1981] vorgeschlagenen und im letzten Abschnitt eingeführten Epipolargeometrie. Hierbei werden korrespondierende Merkmalspunkte verschiedener Ansichten derselben Szene genutzt, um daraus die 3D Bewegung und Struktur zu berechnen. Im Gegensatz zu den direkten Verfahren erfolgt die Zuordnung hier ausschließlich für ausgewählte Bildelemente, die in unterschiedlichen Ansichten der gleichen Szene dasselbe physikalische Objekt abbilden. Diese merkmalsbasierte Korrespondenzanalyse wurde im vorherigen Abschnitt zur Bestimmung der 2D Bildbewegung bereits kurz erläutert.

#### Merkmalsbasierte Korrespondenzanalyse

Die Korrespondenzanalyse wird auf vorher aus dem Bild extrahierten, ausgewählten Bildelementen durchgeführt. Diese Elemente werden durch spezifische Attribute beschrieben. Die Zuordnung geschieht dann anhand dieser Attribute. Je ausgeprägter die Merkmale sind, desto einfacher kann eine eindeutige Zuordnung erfolgen. Nachteilig wirken sich oftmals die zur Extraktion der Merkmale verwendeten Methoden aus, die zusätzliche Unsicherheiten hervorrufen können und Information herausfiltern. Desweiteren können im Allgemeinen keine dichten Merkmalsfelder berechnet werden. Die wesentlichen Vorteile gegenüber intensitäts-

basierten Verfahren sind eine effiziente Realisierbarkeit und weniger Mehrdeutigkeiten, da die Beschreibungskraft der Korrespondenzkandidaten meist höher ist als das Intensitätssignal selbst. Weiterhin sind merkmalsbasierte Verfahren invariant bzgl. der photometrischen Variation bei der Bildgenerierung, da die Merkmale meist Eigenschaften der Szene repräsentieren. Zudem kann die Bestimmung der Verschiebungsvektoren wesentlich genauer erfolgen, da die Bildposition meist subpixelgenau berechnet wird.

Der prinzipielle Ablauf der merkmalsbasierten Korrespondenzanalyse lässt sich in die sequentielle Ausführung folgender Schritte aufteilen: Zunächst müssen markante Bildmerkmale (i) *extrahiert*, danach (ii) möglichst eindeutig *beschrieben* und schließlich (iii) zwischen den einzelnen Ansichten *zugeordnet* werden. Die einzelnen Korrespondenzen werden dabei nur für eine begrenzte Menge diskreter Bildpositionen gesucht, wobei markante Bildstrukturen wie z. B. Ecken zu bevorzugen sind. Diese Form der Unterabtastung des Bildbereichs hat zur Folge, dass homogene Bildregionen, bei denen das Korrespondenzproblem nicht eindeutig gelöst werden kann, prinzipbedingt ausgeschlossen werden. Neben dem klassischen Harris-Eckendetektor [Harris u. Stephens, 1988], bei dem Korrespondenzen zwischen Ecken im Bild durch einen direkten Vergleich von Bildblöcken bestimmt wird, wurden in dieser Arbeit auch verteilungsbasierte Ansätze wie der SIFT- (engl. *SIFT=Scale Invariant Feature Transformation*) [Lowe, 2004] und SURF- (engl. *SURF=Speeded-Up Robust Features*) [Bay u. a., 2008] Deskriptor untersucht.

### Bestimmung der 3D Bewegung

Für eine Menge korrespondierender Merkmalspunkte $\{\mathbf{x}_n^m | n \in \mathcal{K} = \{1,\ldots,M\}\}$ in drei Ansichten räumlich und zeitlich versetzter Kameras beschreibt der sog. *Trifokaltensor* die Beziehung der Punktkorrespondenzen in Form einer geometrischen Bedingungsgleichung. Die trifokale Bedingung kann als eine Erweiterung der Epipolarbedingung um eine weitere Ansicht der Szene in zeitlicher Richtung betrachtet werden. Die durch den Trifokaltensor beschriebene trilineare Bedingung fordert (i) die Einhaltung der Epipolarbedingungen zwischen dem ersten und zweiten Bild nach Rotation und Translation der Kamera, (ii) die Einhaltung der Epipolarbedingungen zwischen dem zweiten und dritten Bild nach erneuter Rotation und Translation der Kamera und (iii) die identische Rekonstruktion der Entfernung aller Punkte von der zweiten Kamera, unabhängig davon, ob die Entfernung aus den ersten oder letzten beiden Bildern rekonstruiert wurde [Stiller u. a., 2009]. So werden u. a. in [Yu u. a., 2006; Dang, 2007; Dang u. Stiller, 2009] Verfahren vorgestellt, die auf Basis des Trifokaltensors eine Schätzung der 3D Eigenbewegung durchführen. Eine ausführliche Untersuchung der trifokalen Bedingung im

Kontext einer kontinuierlichen Selbstkalibrierung von Stereokameras wird in der Arbeit von [Dang, 2007] gegeben. Dabei wird festgestellt, dass Verfahren, die auf der trifokalen Bedingung aufbauen, in der Regel trotz der formalen Vollständigkeit hinter dem sog. Bündelausgleichsverfahren zurückstehen, das Struktur und Bewegung aus ganzen Bildverbänden $\mathcal{G}_T$ schätzt.

Der Begriff Bündelblockausgleichung (engl. *bundle adjustment*), oder auch einfach Bündelausgleich, stammt aus der Photogrammetrie und erlaubt die gleichzeitige Bestimmung der internen und externen Kameraparameter sowie der 3D Struktur der Szene aus mehreren von verschiedenen Orten aus aufgenommenen Bildern [Schmid, 1958; Triggs u. a., 2000; Hartley u. Zisserman, 2004]. Grundlegende Annahme für den Bündelausgleich sind die Starrheit der Szene zwischen den einzelnen Ansichten und die Erfüllung der Kollinearitätsgleichung (engl. *collinearity condition*). Diese besagt, dass ein betrachteter 3D Objektpunkt, sein zugehöriger Bildpunkt und das Projektionszentrum der Kamera auf einer gemeinsamen Gerade liegen müssen. Zielsetzung des Bündelausgleichs ist es nun, durch gleichzeitige Variation der 3D Koordinaten und der Transformationsparameter der einzelnen Ansichten, eine möglichst gute Übereinstimmung zwischen erwarteten und gemessenen Bildpunkten zu erreichen. Da im Falle einer Zentralprojektion jeder Szenenpunkt einen Strahl durch das Projektionszentrum definiert, ergibt sich für die Menge aller Punkte ein Strahlenbündel, welches im Projektionszentrum geschnürt wird. Hieraus ergibt sich der Begriff *Bündel*. Geometrisch kann die Minimierung der Distanz zwischen erwarteten und gemessenen Merkmalspunkten $\{\mathbf{x}_{\mathrm{n}}^{\mathrm{m}} | \mathrm{n} \in \mathcal{K}\}$ und $\{\hat{\mathbf{x}}_{\mathrm{n}}^{\mathrm{m}} | \mathrm{n} \in \mathcal{K}\}$ als Optimierung eines Bündels von Projektionsstrahlen interpretiert werden, welches durch die betrachteten Punkte und das Projektionszentrum der jeweiligen Kameras definiert ist. Für den Fall eines vollständig kalibrierten Stereokamerasystems vereinfacht sich die Schätzung beträchtlich. Im Rahmen dieser Arbeit wurde hierfür u. a. die Verwendung von Einheitsquaternionen [Horn, 1987] zur Registrierung endlicher Punktmengen untersucht [Eisenbeiss, 2007]. Hierbei werden die Transformationsparameter R und $\mathbf{t}$ gesucht, die den mittleren Abstand zwischen transformierter Punktmenge und dazugehörigen korrespondierenden Punkten minimiert. Weiterhin können die Bewegungsparameter auch durch schritthaltende Filterverfahren bestimmt werden [Nister u. a., 2004; Zhu u. a., 2006; Howard, 2008; Cumani u. Guiducci, 2008]. In der Literatur sind solche Ansätze unter dem Begriff *visuelle Odometrie* (engl. *visual odometry*) bekannt. Als Schätzvorschrift ergibt sich

$$(\mathrm{R}', \mathbf{t}') = \underset{\mathrm{R}_t, \mathbf{t}_t}{\arg\min} \left\{ \sum_{t \in \mathcal{T}, \mathrm{n} \in \mathcal{K}} \| \hat{\mathbf{x}}_{\mathrm{n},t}^{\mathrm{m}} - \mathbf{x}_{\mathrm{n},t}^{\mathrm{m}}(\mathrm{R}_t, \mathbf{t}_t) \| \right\}. \tag{2.16}$$

### 2.2.2 Szenenrekonstruktion

Für die dichte Szenenrekonstruktion bei einem rektifizierten Stereokamerasystem existieren hinsichtlich der Repräsentation der Tiefendaten eine Vielzahl unterschiedlicher Möglichkeiten, auf die an dieser Stelle nicht näher eingegangen werden soll. Die Artikel von [Scharstein u. Szeliski, 2001; Brown u. a., 2003] geben hierzu eine Übersicht aktueller Verfahren und Modelle. Eine in diesem Zusammenhang in der Literatur oftmals verwendete Darstellungsform ist die Nutzung einer sog. *Disparitätsfunktion* $\Delta(x,y,t)$. Diese ordnet jedem Bildpunkt $\mathbf{x}_t = (x,y)^{\mathrm{T}}$ in einem vorher definierten Referenzbild einen skalaren Wert zu, mit dem die entsprechende Szenentiefe bestimmt werden kann. Die Zuordnung geschieht im sog. Disparitätsraum $\mathcal{P} = (x,y,\Delta(x,y,t))$ und kann kontinuierlich oder diskret erfolgen. Da die vertikale Koordinate $y$ in beiden Kamerabildern identisch ist, ergibt sich für die Disparitätsfunktion $\Delta(x,y,t) = (x_{\mathrm{l}} - x_{\mathrm{r}})$, welche im Allgemeinen in der Einheit Pixel angegeben wird. Mit Hilfe von (2.7) lässt sich daraus die 3D Rekonstruktion eines Szenenpunktes folgendermaßen berechnen

$$\begin{aligned} \Delta(x,y,t) &= \left(\frac{\mathrm{f}X_{\mathrm{r},t}}{Z_t} + \frac{\mathrm{fb}}{Z_t}\right) - \frac{\mathrm{f}X_{\mathrm{r}}}{Z_t} = \frac{\mathrm{fb}}{Z_t} \\ &\Rightarrow \frac{Z_t}{\mathrm{f}} = \frac{\mathrm{b}}{\Delta(x,y,t)} \Leftrightarrow Z = \frac{\mathrm{bf}}{\Delta(x,y,t)}\,. \end{aligned} \tag{2.17}$$

Mit Ausnahme von Bereichen, die in der jeweils anderen Ansicht verdeckt werden, kann somit bei Kenntnis der wahren Konfiguration $\Delta_t \in \mathcal{P}$ theoretisch für jeden abgebildeten Szenenpunkt die zugehörige Tiefe berechnet werden. $\Delta_t$ wird im Weiteren auch als Disparitätskarte bezeichnet. Alternativ dazu kann die Rekonstruktion der Szene auch im 3D Strukturraum $\mathcal{Z}$ erfolgen, wie u. a. in [Larsen u. a., 2006] beschrieben. Ein wesentlicher Vorteil hierbei ist die Definition des Rekonstruktionsmodells im 3D Raum, womit Bedingungen an das Schätzergebnis direkt an die gesuchte Szenenstruktur gestellt werden können.

Das Ergebnis der stereoskopischen Rekonstruktion wird durch $\mathbf{Z}_t$ beschrieben. Die Suche nach einer optimalen Konfiguration von $\mathbf{Z}_t$ auf der Basis rektifizierter Stereobildpaare wird in dieser Arbeit als Labelingproblem formuliert und effizient mit Hilfe diskreter Optimierungsverfahren gelöst.

## 2.3 Das Labelingproblem

Ein Labelingproblem (engl. *label=Beschriftung, Etikett, Kennzeichen*) ist spezifiziert durch die (nicht leere) Indexmenge $\mathcal{I}$ und eine Menge von Labelwerten

$\mathcal{L} = \{1, \ldots, j, \ldots, J\}$. Die Definitionen können hierbei gleichermaßen im diskreten als auch im kontinuierlichen Wertebereich erfolgen. In dieser Arbeit repräsentieren die Variablen in $\mathcal{I}$ die Bildpunktorte, welche regelmäßig auf dem 2D Bildgitter verteilt sind. Aus Gründen der einfacheren Notation wird $\mathcal{I} = \{1, 2, \ldots, N\}$ gewählt, wobei $N = U \cdot V$ die Anzahl der Bildpunkte ist. Diese Notation wird auch bei der Beschreibung mehrdimensionaler Zufallsfelder verwendet, die im nächsten Abschnitt für die stochastische Modellierung des Segmentierprozesses eingeführt werden.

Das Labelingproblem besteht nun in der Zuordnung eines Labelwertes aus $\mathcal{L}$ zu jeder Variable in $\mathcal{I}$. Das Ergebnis dieser Zuweisung wird als Labeling $\mathbf{l} = (l_1, \ldots, l_n, \ldots, l_N)$ bezeichnet und drückt eine Abbildung (engl. *mapping*) $\mathbf{l} : \mathcal{I} \mapsto \mathcal{L}$ aus, wie in Abbildung 2.3(a) skizziert. Ist für alle Bildpunkte in $\mathcal{I}$ der Wertebereich in $\mathcal{L}$ identisch, ergibt sich die Menge aller möglichen Labelkonfigurationen zu $\mathcal{F} = \mathcal{L} \times \mathcal{L} \times \ldots \times \mathcal{L} = \mathcal{L}^N$. Für die Lösung des Labelproblems bei einer Menge von $N$ Bildpunkten und $J$ möglichen Labelwerten, muss die optimale Labelkonfiguration aus einer Menge von insgesamt $J^N$ möglichen Konfigurationen in $\mathcal{F}$ gefunden werden.

Bei der in dieser Arbeit angestrebten dichten Szenensegmentierung aus Stereobildfolgen existieren insgesamt zwei, getrennt voneinander zu betrachtende, Labeldefinitionen für $\mathcal{L}$. Das erste Labeling repräsentiert das Ergebnis der dichten Szenenrekonstruktion, wie in Abschnitt 2.2.2 beschrieben. Das zweite Labeling drückt die Szenensegmentierung selbst aus. Hierbei sollen Gruppierungen von ähnlich bewegten Szenenpunkten aus der Bildsequenz herausgelöst und zusammenfassend durch jeweils einen spezifischen Labelwert $j \in \mathcal{L}$ beschrieben werden. Das der Segmentierung zugrunde liegende Verschiebungsfeld wird hierbei segmentweise durch ein noch zu spezifizierendes Bewegungsmodell beschrieben. Eine optimale Labelkonfiguration zerlegt das Bild in sich nicht überlappende Bildsegmente $\mathcal{X}^1, \mathcal{X}^2, \ldots, \mathcal{X}^J \subset \mathcal{X}$. Diese Zerlegung ist die in dieser Arbeit angestrebte Szenensegmentierung.

## 2.4 Stochastische Modellierung

Die Modellierung der einzelnen Größen der Szenensegmentierung erfolgt in der Form eines stochastischen Prozesses $\mathbf{X}_\mathcal{I} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$, der aus einer endlichen Menge von Zufallsvariablen $\mathbf{X}_{n \in \mathcal{I}}$ besteht. Durch die gemeinsame Auswertung der in $\mathbf{X}_\mathcal{I}$ enthaltenen Zufallsgrößen kann die örtliche und zeitliche Entwicklung eines Vorgangs der realen Welt modellhaft abgebildet und die starken stochastischen Bindungen natürlicher Prozesse können formal gefasst werden. Für das so defi-

nierte Zufallsfeld (engl. *random field*) bezeichnet $\mathbf{x}_\mathcal{I} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ eine spezielle Realisation des Feldes. Dabei ist der Zustandsraum einer einzelnen Variable definiert mit $\omega_n$, d. h. $\mathbf{X}_n \in \omega_n, \forall n \in \mathcal{I}$. Der Zustandsraum des gesamten Zufallsfeldes ergibt sich somit zu $\Omega_\mathcal{I} = {}^\times\prod_{n\in\mathcal{I}} \omega_n$. Besteht der Zustandsraum aus einer abzählbaren Menge, so handelt es sich um einen kontinuierlichen Prozess. Ist $\omega$ ein Intervall, so ist es ein stetiger Prozess. Zu einem stochastischen Prozess wird $\mathbf{X}_\mathcal{I}$ durch die Einführung des *Wahrscheinlichkeitsmaßes P*, mit welchem die Wahrscheinlichkeit $P(\mathbf{X}_\mathcal{I} = \mathbf{x}_\mathcal{I})$ bestimmter Realisationen des Prozesses quantifiziert werden kann. Für den allgemeinen Fall wird dieses Maß als Verteilungsdichte (engl. *probability distribution*) des Zufallsfeldes bezeichnet. Handelt es sich bei $\mathbf{X}_\mathcal{I}$ um eine diskrete Größe, spricht man von einer Verteilungsfunktion (engl. *probability mass function*). Die Auswertung dieses stochastischen Modells hin zu einer gemeinsamen Verbundverteilung stellt schon für kleine Zufallsfelder ein komplexes Problem dar, da hierfür die bedingte Verteilung $P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{\mathcal{I}_{\setminus n}} = \mathbf{x}_{\mathcal{I}_{\setminus n}})$ jeder einzelnen Variablen $\mathbf{X}_n$ in Abhängigkeit aller anderen Variablen bestimmt werden muss. $\mathcal{I}_{\setminus n}$ symbolisiert hier die Indexmenge $\{1, \ldots, n-1, n+1, \ldots, N\}$ ohne Element $n$. Durch die Ausnutzung der Markov-Eigenschaft, welche im Folgenden näher beschrieben wird, kann das Modell jedoch wesentlich vereinfacht werden.

### 2.4.1 Markov-Ketten

Die Komplexität des Modells kann deutlich reduziert werden, falls die Indexmenge $\mathcal{I}$ einer ausgezeichneten Ordnungsrelation unterliegt. Oftmals kann hierbei eine kausale Abhängigkeit der Elemente in $\mathcal{I}$ genutzt werden, um die Verbundverteilung des Zufallsfeldes als Produkt der univariaten bedingten Wahrscheinlichkeiten seiner Komponenten zu beschreiben. Unter Anwendung der Kettenregel ergibt sich dann folgender Ausdruck:

$$P(\mathbf{X}_\mathcal{I} = \mathbf{x}_\mathcal{I}) = \prod_{n\in\mathcal{I}} P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{\mathcal{I}_{<n}} = \mathbf{x}_{\mathcal{I}_{<n}}) \,. \tag{2.18}$$

Die Menge $\mathcal{I}_{<n}$ beinhaltet hierbei die Vorgänger von $\mathbf{X}_n$. $\mathbf{X}_\mathcal{I}$ bildet genau dann eine Markov-Kette $u$-ter Ordnung, falls

$$\begin{aligned} &P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{\mathcal{I}_{<n}} = \mathbf{x}_{\mathcal{I}_{<n}}) = \\ &P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_v = \mathbf{x}_v, \forall v \in \mathcal{I} \text{ mit } (n-u) \leq v < n) \quad \forall n \in \mathcal{I} \,. \end{aligned} \tag{2.19}$$

Diese Forderung wird auch als *einseitige Markov-Eigenschaft* $u$-ter Ordnung des Zufallsprozesses bezeichnet. Die bedingte Verteilung des Zustandes $\mathbf{X}_n$ kann also vollständig durch die Kenntnis der Realisationen der $u$ direkten Vorgänger be-

stimmt werden. Dies wird auch als Zufallsprozess mit endlichem Gedächtnis bezeichnet. Von besonderem Interesse sind Markov-Ketten erster Ordnung mit $u = 1$, also Ketten die nur die direkten Nachbarn auswerten:

$$P(\mathbf{X}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}}) = P(\mathbf{X}_0 = \mathbf{x}_0) \prod_{n=2}^{N} P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{n-1} = \mathbf{x}_{n-1}) \,. \tag{2.20}$$

Die bedingte Verteilungsdichte $P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{n-1} = \mathbf{x}_{n-1})$ wird auch als Übergangsdichte des Markov-Prozesses bezeichnet. Aufgrund der Markov-Eigenschaft lässt sich somit jede gemeinsame Verteilungsdichte durch eine Anfangsverteilungsdichte $p(\mathbf{X}_0 = \mathbf{x}_0)$ und die Übergangsdichten darstellen. Für homogene Markov-Ketten erfüllen die Übergangsdichten die sog. Chapman-Kolmogoroff-Gleichung [Chapman, 1928; Kolmogoroff, 1933]

$$p(\mathbf{x}_3 | \mathbf{x}_1) = \int_{-\infty}^{\infty} p(\mathbf{x}_3 | \mathbf{x}_2) p(\mathbf{x}_2 | \mathbf{x}_1) d\mathbf{x}_2 \,. \tag{2.21}$$

Diese Struktur formuliert die Grundlage der rekursiven Schätzverfahren zur Bestimmung zeitveränderlicher Modellparameter, welche auch in dieser Arbeit zum Einsatz kommen.

Aufgrund der fehlenden Ordnungsrelation bei der stochastischen Modellierung von Bildern ist eine einseitige Entwicklung, wie bei (2.19), jedoch nur eingeschränkt möglich. Hierfür wird die *beidseitige Markov-Eigenschaft* $u$-ter Ordnung definiert, welche erfüllt ist, falls

$$\begin{aligned} &P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{\mathcal{I}_{\setminus n}} = \mathbf{x}_{\mathcal{I}_{\setminus n}}) = \\ &P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_v = \mathbf{x}_v, \forall v \in \mathcal{I}_{\setminus n}, \text{ mit } 0 < |v - n| \leq u) \quad \forall n \in \mathcal{I} \,. \end{aligned} \tag{2.22}$$

Die Gleichung besagt, dass die Wahrscheinlichkeit für die Belegung der Zufallsvariablen $\mathbf{X}_n$ unter Berücksichtigung der Werte aller anderen Zufallsvariablen gleich der bedingten Wahrscheinlichkeit lediglich unter Berücksichtigung der Werte in der lokalen Umgebung ist. Diese Verteilung wird aufgrund ihrer Begrenztheit auf umliegende Zufallsvariablen auch als *lokale Charakteristik* bezeichnet. Im Kontext mehrdimensionaler Zufallsfelder wird in diesem Zusammenhang der Begriff der Nachbarschaft, bzw. des Nachbarschaftsystems in Abschnitt 2.4.2 eingeführt.

Mit den oben getroffenen Annahmen ergeben sich somit für einen Zufallsprozess die im Folgenden nochmals zusammengefassten Eigenschaften.

**Definition 2.1 (Markov-Prozess)**
*Ein Zufallsprozess $\mathbf{X}_{\mathcal{I}}$ ist ein Markov-Prozess, falls*

*(i) (Positivität)*
*für alle Verbundverteilungen gilt: $P(\mathbf{X}_{\mathcal{I}_{\setminus n}} = \mathbf{x}_{\mathcal{I}_{\setminus n}}) > 0, \forall \mathbf{x}_{\mathcal{I}_{\setminus n}}, \Omega_{\mathcal{I}_{\setminus n}}, n \in \mathcal{I}$. Als Konsequenz hieraus existieren alle bedingten Verteilungen $P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{\mathcal{I}_{\setminus n}} = \mathbf{x}_{\mathcal{I}_{\setminus n}})$ des Zufallsfeldes, womit auch $P(\mathbf{X}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}}) > 0 \, \forall \mathbf{x}_{\mathcal{I}} \in \Omega_{\mathcal{I}}$ erfüllt ist. Nach [Li, 2009] kann dies in der Praxis als erfüllt betrachtet werden.*

*(ii) (Markov-Eigenschaft)*
*die in* (2.19) *und in* (2.22) *beschriebene einseitige bzw. beidseitige Markov-Eigenschaft für alle Elemente $\mathbf{X}_n, n \in \mathcal{I}$ erfüllt ist.*

Die Erfüllung dieser Eigenschaften wird im weiteren Verlauf der Arbeit vorausgesetzt und für die Auswertung örtlich und zeitlich abhängiger Zufallsprozesse eingesetzt. Da sich jedoch mit Markov-Ketten die 2D Struktur von Bilddaten nur in sehr eingeschränktem Maße modellieren lässt, ist es notwendig, die Markov-Eigenschaft auf mehrdimensionale Zufallsfelder zu erweitern, wie im Folgenden gezeigt wird.

### 2.4.2 Gibbs/Markov-Felder

Der Bildinhalt wird in diesem Zusammenhang als Realisierung einer mehrdimensionalen Zufallsvariable betrachtet. Diese Zufallsvariable besitzt eine bestimmte, die Struktur des gesamten Bildes beschreibende, Verteilung. Gerade die Eigenschaft der ganzheitlichen Beschreibung der Bildinhalte ist es, die Zufallsfelder in der statistischen Bildanalyse so populär machen. Da es im Fall eines mehrdimensionalen Zufallsfeldes, das über dem 2D Bildgitter definiert ist, keine bevorzugte Richtung für eine der Markov-Kette entsprechenden Kausalität gibt, ist die Erweiterung des zuvor getroffenen Abhängigkeitsgedankens zufallsbehafteter Größen jedoch nicht trivial.

#### Markov-Felder

Die Idee, mit dem einseitigen Abhängigkeitsgedanken auch seine Kausalität aus dem Eindimensionalen in höhere Dimensionen zu übertragen, führen u. a. zu den

sog. *Markov Mesh Random Fields* [Abend u. a., 1965; Kanal, 1980]. Bei diesem Modell besteht das Zufallsfeld aus den Bildpunkten, wobei die Bildpunktorte in festgelegter Ordnung die Indexmenge $\mathcal{I}$ bilden. Eine Realisation des so definierten Zufallsfeldes wird darin als Markov-Kette modelliert, deren einseitige Abhängigkeitsmengen bestimmte Eigenschaften – identisch der in (2.19) definierten Ordnungsrelation – erfüllen. So definierte Vorgängermengen bzgl. einer Zufallsvariablen können als eine Art „nächster Nachbar"-Modell mit spezieller Ausbreitungsrichtung interpretiert werden. Die „Nachbarschaft" ist hierbei an gewisse Konsistenzbedingungen geknüpft, die jedoch aufgrund der 1D Struktur des Modells die stochastischen Bindungen realer Bilddaten nur sehr begrenzt fassen.

Alternativ dazu kann die oben eingeführte beidseitige Markov-Eigenschaft auf mehrdimensionale Zufallsfelder verallgemeinert werden. Analog zu den Ausführungen einer Markov-Kette, muss hierfür zunächst für jede Zufallsvariable eine mehrdimensionale Abhängigkeitsmenge definiert werden. Dazu wird der bereits erwähnte Begriff eines Nachbarschaftsystems $\mathcal{N}_\mathcal{I}$ auf $\mathcal{I}$ nun näher spezifiziert.

**Definition 2.2 (Nachbarschaftssystem)**
*Ein Nachbarschaftssystem $\mathcal{N}_\mathcal{I} = \{\mathcal{N}_n | n \in \mathcal{I}\}$ innerhalb eines Zufallsfeldes ist nach [Li, 2009] durch zwei grundsätzliche Eigenschaften gekennzeichnet:*

*(i) (Irreflexivität)*
*Eine Zufallsvariable n ist nicht Element ihrer eigenen Nachbarschaft:*
*$\mathcal{N}_n$: $n \notin \mathcal{N}_n \ \forall n \in \mathcal{I}$*

*(ii) (Symmetrie)*
*Wenn n ein Element einer Nachbarschaft von u ist, so ist u ein Element der Nachbarschaft von n: $n \in \mathcal{N}_u \Leftrightarrow u \in \mathcal{N}_n \ \forall n, u \in \mathcal{I}$*

Das Nachbarschaftssystem $\mathcal{N}_\mathcal{I}$ und die Entitäten $\mathcal{I}$ konstituieren einen ungerichteten Graphen $\mathrm{G}(\mathcal{I}, \mathcal{N}_\mathcal{I})$. Die Knoten des Graphen entsprechen der Indexmenge, die Kanten der jeweiligen Nachbarschaftsrelation. Für die Konstruktion eines Nachbarschaftssystems werden Nachbarschaftsrelationen meist anhand einer Distanz $r$, z. B. der euklidische Abstand, zur zentralen Zufallsvariablen $n$ erstellt:

$$\mathcal{N}_n = \{u \in \mathcal{I} \mid \|\mathbf{x}_n - \mathbf{x}_u\| < r \,,\, u \neq n\} \tag{2.23}$$

Es gilt zu beachten, dass Nachbarschaften unter Einhaltung der oben aufgeführten Bedingungen völlig beliebig definiert werden können. Eine Ordnungsrelation auf $\mathcal{I}$ ist somit nicht erforderlich. Für viele Anwendungen in der Bildverarbeitung erscheint jedoch die Wahl eines regelmäßigen Nachbarschaftssystems als sinnvoll.

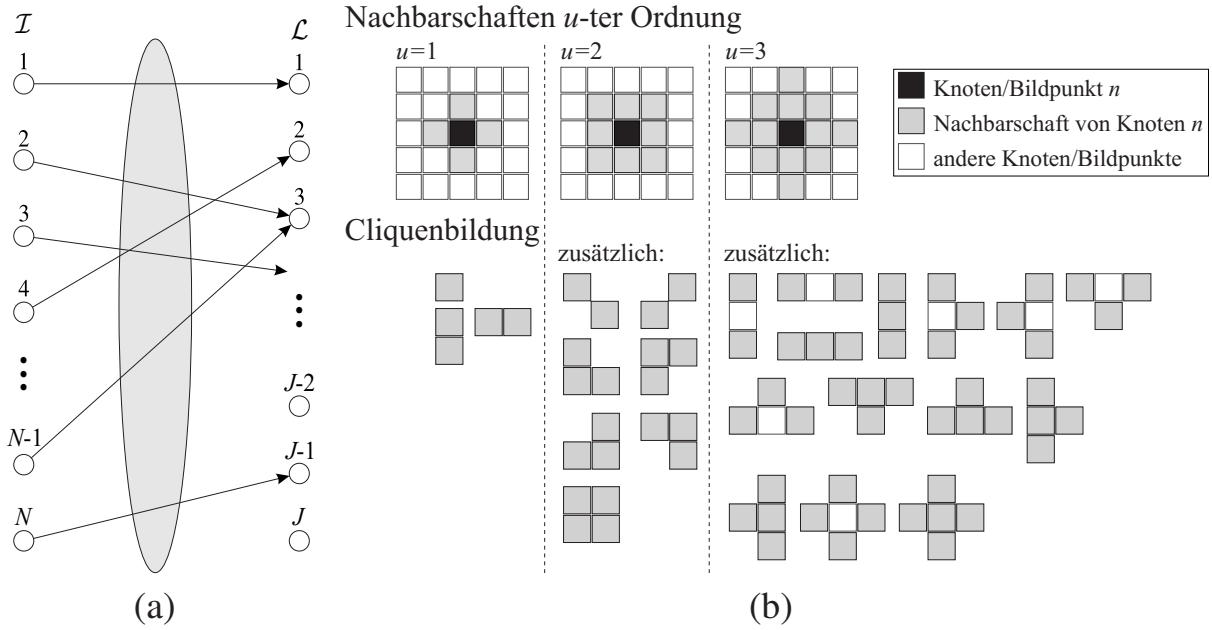



**Abbildung 2.3**: (a) [Angelehnt an [Li, 2009]] Ein Labeling kann als Abbildung der Punktmenge $\mathcal{I}$ auf den Satz von Labelwerten $\mathcal{L}$ betrachtet werden. (b) Beispiele für Nachbarschaftssysteme $\mathcal{N}_{\mathcal{I}}$ und entsprechender Cliquenbildung $\mathcal{C}_1, \mathcal{C}_2, \ldots$ unterschiedlicher Ordnung $u$ in einem regelmäßigen Gitter.

Die Nachbarschaft besteht hier aus den $u$ nächsten Gitterpunkten relativ zum Referenzpunkt. Zufallsvariablen, die Punkte am Bildrand repräsentieren, müssen hier gesondert behandelt werden.

**Definition 2.3 (Markov-Feld)**
*Ein Zufallsfeld $\mathbf{X}_{\mathcal{I}}$ ist nun ein Markov-Feld (engl. Markov random field) (MRF), wenn folgender Ausdruck gilt:*

$$P(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{\mathcal{I}\setminus n} = \mathbf{x}_{\mathcal{I}\setminus n}) = P(\mathbf{X}_n | \mathbf{X}_u, \forall u \in \mathcal{N}_n) \qquad \forall n \in \mathcal{I} \tag{2.24}$$

Für den eindimensionalen Fall geht das Markov-Feld in eine Markov-Kette über, wie aus (2.19) klar wird. Die Gleichung drückt die Tatsache aus, dass die bedingte Wahrscheinlichkeit der Variable $\mathbf{X}_n$ ausschließlich von seiner Nachbarschaft $\mathcal{N}_n$ abhängig ist[10]. Diese prinzipiell wünschenswerte Eigenschaft stößt jedoch in der Praxis auf erhebliche Schwierigkeiten bei der Anwendung, welche von [Derin, 1986; Li, 2009] wie folgt zusammengefasst werden:

[10]Eine weitere, sehr anschauliche Definition eines MRF, welches durch die Einführung des Begriffs der Umrandung eine Art probabilistische Abschirmung für einzelne Gebiete definiert, kann u. a. in der Arbeit von [Stiller, 1994] nachgelesen werden.

- ◇ Aufgrund der meist fehlenden kausalen Abhängigkeiten eines mehrdimensionalen Zufallsfeldes, gestaltet sich die Ermittlung der Verbundwahrscheilichkeit auf Basis der lokalen Charakteristik[11] des Zufallsprozesses ungleich schwerer als bei der oben vorgestellten Markov-Kette.
- ◇ Die gemeinsame Verteilung des Zufallsprozesses aus der lokalen Charakteristik ist nicht ohne Weiteres ersichtlich.
- ◇ Die gegebenen lokalen Charakteristiken definieren nicht notwendigerweise konsistent ein Zufallsfeld mit einer gemeinsamen Verteilung der Komponenten.

Aufgrund dieser Tatsachen fanden Markov-Zufallsfelder bis in die siebziger Jahre hinein, trotz ihrer einfachen und theoretisch eleganten Struktur zur Beschreibung stochastischer Abhängigkeiten, nur sehr eingeschränkt Anwendung. Der praktische Nutzen steigerte sich jedoch drastisch mit dem bekannt werden des *Hammersley-Clifford-Theorems*. Dieser von [Hammersley u. Clifford, 1971] eingeführte und später von [Besag, 1974] vereinfachte Beweis der Äquivalenz eines MRF und einer Gibbs-Verteilung erlaubt die explizite Formulierung der Verbundverteilung eines MRF und somit eine numerisch handhabbare Möglichkeit der Auswertung mehrdimensionaler Markov-Felder.

### Gibbs-Felder

Bei Gibbs-Feldern handelt es sich um eine, durch die Form der gemeinsamen Verteilung der Komponenten definierten Klasse von Zufallsfeldern, deren Verbundwahrscheinlichkeitsdichte durch die spezielle Struktur einer Gibbs-Verteilung charakterisiert ist. Da diese Charakterisierung einer Analogie zur statistischen Physik entspringt, sind auch die meisten Bezeichnungen von dort übernommen. Betrachtet wird in diesem Zusammenhang eine Konfiguration mehrerer Teilchen (etwa Atome oder Moleküle), die bestimmte Zustände einnehmen können. Die auf ein einzelnes Teilchen wirkenden Kräfte werden beschrieben durch ein äußeres Kraftfeld. Die Kräftebilanz wird um die Kräfte zwischen einzelnen Teilchen erweitert. Da zu jeder Kraft eine entsprechende Gegenkraft wirkt, bedingen die im System vorhandenen Wechselwirkungen eine symmetrische Nachbarschaftsrelation auf den Teilchen. Überträgt man nun dieses Modell auf das Bildgitter, wobei die Teilchen die einzelnen Bildpunkte darstellen, werden in Analogie zu dem physikalischen Teilchenmodell folgende Begriffe zur näheren Beschreibung der Gibbs-Verteilung eingeführt.

[11] Hängt die lokale Charakteristik nur von der Konfiguration des Nachbarschaftssystems, jedoch nicht von der relativen Lage auf dem Bildgitter ab, spricht man von einem „homogenen" MRF.

**Definition 2.4 (Clique)**
*Sei $\mathcal{N}_\mathcal{I}$ ein Nachbarschaftssystem über dem Bildgitter $\mathcal{R}$. Eine Clique ist dann eine (nicht leere) Teilmenge der Entitäten $\mathcal{I}$ des ungerichteten Graphen $\mathrm{G}(\mathcal{I},\mathcal{N}_\mathcal{I})$ für die gilt, dass alle Indexpaare $c$ der Teilmenge benachbart sind. Sie besteht entweder aus einer Einerclique $c_1 = \langle n \rangle$, die der betrachteten Entität selbst entspricht, oder einem Paar $c_2 = \langle n,u \rangle$, Tripel $c_3 = \langle n,u,v \rangle$, usw. benachbarter Entitäten*[12]*. Entsprechend setzt sich dies für höherwertige Cliquen fort. Die Menge aller Cliquen $\mathcal{C}_1 = \{\langle n \rangle \,|\, n \in \mathcal{I}\}$, $\mathcal{C}_2 = \{\langle n,u \rangle \,|\, u \in \mathcal{N}_n, n \in \mathcal{I}\}$, $\cdots$ wird mit $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3 \cdots$ bezeichnet.*

Es ist unschwer zu erkennen, dass die Definition einer Clique sehr eng mit dem Begriff der Nachbarschaftssystems verbunden ist. Bei der Wahl des Nachbarschaftssystems sollte beachtet werden, dass die Anzahl der Cliquen mit steigender Ordnung $u$ stark zunimmt. Abbildung 2.3(b) zeigt einige gebräuchliche Nachbarschaftssysteme unterschiedlicher Ordnung $u$ und entsprechende Cliquenbildung.

Da innerhalb einer Clique alle Elemente paarweise benachbart sind, ergibt sich nach Freischneiden[13] des physikalischen Teilchenmodells ein entsprechendes Potential. Bei einelementigen Cliquen ist dieses Potential durch das äußere Kraftfeld bestimmt.

**Definition 2.5 (Cliquenpotential)**
*Ein Potential $V$ eines Feldes $\mathbf{X}_\mathcal{I}$ ist eine Menge von Einzelpotentialen $V_c$, mit $c \subset \mathcal{C}$. Wie durch die Indexierung bereits angedeutet, sind diese Einzelpotentiale durch die oben eingeführten Cliquen definiert, welche die Bedingung $V_c = 0 \;\forall\, c \notin \mathcal{C}$ erfüllen.*

Mit dem so definierten Cliquenpotential kann die Energie des Systems aller Teilchen dann einfach als Summe aller Einzelpotentiale ausgedrückt werden:

$$H(\mathbf{X}_\mathcal{I}) = \sum_{c \in \mathcal{C}} V_c \,. \tag{2.25}$$

Dieses Maß wird in der statistischen Mechanik als Hamiltonfunktion bezeichnet und hängt indirekt über $V_c$ von der jeweiligen Cliquenkonfiguration ab.

[12] Es muss beachtet werden, dass die Einträge einer Clique geordnet sind. So drücken z. B. $\langle n,u \rangle$ und $\langle u,n \rangle$ unterschiedliche Cliquen aus.

[13] D.h. dem Aufsummieren aller Wechselwirkungskräfte nach der Kirchhoff'schen Knotenregel.

**Definition 2.6 (Gibbs-Feld)**
*Ein Zufallsfeld ist nun ein Gibbs-Feld bzgl. einer speziellen Cliquenkonfiguration $\mathcal{C}$, wenn $\mathbf{X}_{\mathcal{I}}$ der Gibbs-Verteilung*

$$P(\mathbf{X}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}}) = Z^{-1} \exp\left[-H(\mathbf{x}_{\mathcal{I}})\right], \text{ mit}$$
$$Z = \sum_{\mathbf{x}_{\mathcal{I}} \in \Omega_{\mathcal{I}}} \exp\left[-H(\mathbf{x}_{\mathcal{I}})\right] \in \mathbb{R}_+ \tag{2.26}$$

*entspricht.*

Ein Gibbs-Feld wird als homogen bezeichnet, falls $V_c$ unabhängig von der relativen Position von $c$ in $\mathcal{I}$ ist. Aufgrund der besseren mathematischen Handhabbarkeit wird – wie bei den meisten in der Praxis verwendeten Modellen – auch in der vorliegenden Arbeit die Annahme der Homogenität getroffen. Z ist eine Normierungskonstante, die aus obigem Ausdruck ein Wahrscheinlichkeitsmaß macht und als Zustandssumme (engl. *partition function*) bezeichnet wird. In der praktischen Anwendung entzieht sich diese Größe jedoch oftmals einer analytischen oder numerischen Bestimmung. Zur Lösung des Ausdrucks können approximative Lösungen genutzt werden, auf die im Weiteren noch näher eingegangen wird. Bemerkenswert ist, dass jede Verteilung eines Zufallsfeldes in dieser Form dargestellt werden kann, was sich auch in dem physikalischen Sachverhalt unmittelbar widerspiegelt, dass die Konfiguration eines beliebigen natürlichen Prozesses umso wahrscheinlicher ist, je kleiner seine Energie wird.

Mit Hilfe von Gibbs-Zufallsfeldern ist es nun möglich, die globalen Eigenschaften von $\mathbf{X}_{\mathcal{I}}$ in Form einer gemeinsamen Verbundverteilung konsistent auszudrücken. Diese Überlegenheit gegenüber der Beschreibung mit bedingten Wahrscheinlichkeiten eines MRF veranlasste [Whittle, 1963] sogar zur Aussage, Zufallsfelder grundsätzlich durch Gibbs-Felder zu modellieren. Dem steht die intuitivere, da die lokalen Eigenschaften ausdrückende, Beschreibung eines MRF gegenüber. Die Frage nach dem „besseren" Modell ist seit dem Nachweis der Äquivalenz von Markov-Feldern und Gibbs-Feldern u. a. durch [Hammersley u. Clifford, 1971; Spitzer, 1971] hinfällig. Als Konsequenz hieraus ist es also mit der Gibbs-Verteilung möglich, die Vorteile einer Modellierung mit Hilfe von Markov-Feldern zu nutzen und gleichzeitig die dadurch entstehenden Einschränkungen hinsichtlich einer global konsistenten Formulierung zu kompensieren.

### Gibbs-Markov-Äquivalenz

In den letzten Abschnitten wurden Markov-Felder über die lokale Markov-Bedingung charakterisiert, während Gibbs-Felder hingegen über die Verbund-

wahrscheinlichkeit in Form der Gibbs-Verteilung beschrieben wurde. Hier soll die enge Beziehung von Markov- und Gibbs-Feldern vorgestellt werden. Wie bereits erwähnt, kann dies mit Hilfe des Hammersley-Clifford-Theorems gezeigt werden, welches die Gleichheit dieser beiden Beschreibungen beweist. Mit Referenz auf [Grimmett, 1973; Griffeath, 1976; Li, 2009; Besag, 1974] wird in dieser Arbeit jedoch auf einen detaillierten Beweis dieses Theorems verzichtet. Vor allem wichtig für das weitere Verständnis sind die folgenden, daraus getroffenen Feststellungen:

(i) Jedes Gibbs-Feld $\mathbf{X}_{\mathcal{I}}$ auf $\mathcal{I}$ bezüglich des Nachbarschaftssystems $\mathcal{N}_{\mathcal{I}}$ ist ein Markov-Feld bezüglich desselben Nachbarschaftssystems $\mathcal{N}_{\mathcal{I}}$.

(ii) Ein Zufallsfeld $\mathbf{X}_{\mathcal{I}}$ ist ein Markov-Feld bzgl. des Nachbarschaftssystems $\mathcal{N}_{\mathcal{I}}$ genau dann, wenn $\mathbf{X}_{\mathcal{I}}$ ein Gibbs-Feld bzgl. $\mathcal{N}_{\mathcal{I}}$ ist. Lokale Charakteristik und Cliquenpotentiale lassen sich somit ineinander überführen.

Es lässt sich somit beweisen, dass Gibbs-Verteilungen und Markov-Felder äquivalent sind, womit zu jedem MRF eine passende Beschreibung in Form eines Energiefunktionals existiert. Oftmals werden für Anwendungen in der Bildverarbeitung Energiefunktionale verwendet, die sich auf ein- und zweielementige Cliquen beschränken, also

$$H(\mathbf{X}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}}) = \sum_{n \in \mathcal{I}} \mathrm{V}_{\langle n \rangle}(\mathbf{x}_{\mathcal{I}}) + \frac{1}{2} \sum_{n \in \mathcal{I}} \sum_{u \in \mathcal{N}_n} \mathrm{V}_{\langle n,u \rangle}(\mathbf{x}_{\mathcal{I}}) \,. \tag{2.27}$$

Die Ausdrücke „$\sum_{\langle n \rangle \in \mathcal{I}}$" und „$\sum_{\langle n \rangle \in \mathcal{C}_1}$" bzw. „$\sum_{n \in \mathcal{I}} \sum_{u \in \mathcal{N}_n}$" und „$\sum_{\langle n,u \rangle \in \mathcal{C}_2}$" sind hierbei identisch und können analog benutzt werden. Für Zufallsfelder mit einer Verteilung, die durch

$$H(\mathbf{X}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}}) = \sum_{n \in \mathcal{I}} \mathbf{x}_n G(\mathbf{x}_n) + \sum_{n \in \mathcal{I}} \sum_{u \in \mathcal{N}_n} \lambda_{\langle n,u \rangle} \mathbf{x}_n \mathbf{x}_u \,, \tag{2.28}$$

bestimmt ist, wurde von [Besag, 1974] der Begriff *auto-Modelle* geprägt. Handelt es sich bei $\mathbf{X}_{\mathcal{I}}$ um eine diskrete Variable, spricht man von *auto-logistischen Modellen*. $\lambda_{\langle n,u \rangle}$ drückt hier ein Regularisierungskonstante aus, die den Einfluss der Nachbarschaftsrelationen auf die Gesamtenergie steuert. $G$ stellt eine beliebige Funktion dar. Dieses sehr einfache Modell berücksichtigt offensichtlich nur Wechselwirkungen zwischen jeweils höchstens zwei Komponenten innerhalb der Nachbarschaftsumgebung. Als Spezialfall sind hier das Ising-Modell [Ising, 1925]

$$H(\mathbf{X}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}}) = - \sum_{\langle n,u \rangle \in \mathcal{C}_2} \lambda_{\langle n,u \rangle} \mathbf{x}_n \mathbf{x}_u \tag{2.29}$$

und das Potts-Modell [Geman u. Geman, 1984]

$$H(\mathbf{X}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}}) = -\sum_{\langle n,u\rangle \in \mathcal{C}_2} \lambda_{\langle n,u\rangle} \delta_{\mathbf{x}_n,\mathbf{x}_u}, \text{ mit } \delta_{\mathbf{x}_n,\mathbf{x}_u} = \begin{cases} 0 & \text{falls } \mathbf{x}_n \neq \mathbf{x}_u \\ 1 & \text{falls } \mathbf{x}_n = \mathbf{x}_u \end{cases} \quad (2.30)$$

zu nennen. Letzteres kann als Verallgemeinerung des speziellen Ising-Modells in (2.29) angesehen werden, da es nicht auf binäre Werte $\omega_n = \{-1, 1\}\ \forall n \in \mathcal{I}$ beschränkt ist, sondern für einen beliebigen Wertebereich verwendet werden kann.

## 2.5 Lösung des globalen Optimierungsproblems

Ist man an der tatsächlichen Verbundverteilung des Zufallsfeldes interessiert, muss das Energiefunktional $H(\mathbf{X}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}})$ vollständig ausgewertet werden. Hierbei wird die Form der Verteilung durch eine spezielle Funktion und einen Satz von Parametern charakterisiert, welche zusammen das Modell der Wahrscheinlichkeitsverteilung definieren. Für den Fall der Verbundwahrscheinlichkeit eines MRF ist dies die Gibbsfunktion selbst und ein Satz von Parametern für die Cliquenpotentiale. Durch das enthaltene Beobachtungsmodell ergeben sich meist weitere Parameter, die zusätzlich bestimmt werden müssen. Wie (2.26) zeigt, ist zur Berechnung der Gibbs-Verteilung die Bestimmung der Zustandssumme Z notwendig. Da es sich dabei um die Summe[14] über alle Konfigurationen in $\mathcal{F}$ handelt, ist die Auswertung dieses Ausdrucks in der Praxis meist nicht bewältigbar. Zur Schätzung der Verbundverteilung wird deshalb in dieser Arbeit ein approximatives Verfahren verwendet, welches in Abschnitt 4.2.2 näher vorgestellt wird.

### 2.5.1 Formulierung des Problems als MAP-Schätzer

Ist man an der Bestimmung nur eines bestimmten Modalwertes der Verteilungsfunktion des Zufallsfeldes interessiert, kann Z vernachlässigt werden, da diese nur als proportionaler Faktor in die Berechnung mit eingeht. Die für die Schätzung der Wahrscheinlichkeit einer Realisation $\mathbf{X}_{\mathcal{I}} = \mathbf{x}_{\mathcal{I}}$ meist zusätzlich genutzte Information einer beobachteten Größe $\mathbf{y}_{\mathcal{I}}$, wird dabei mit $\mathbf{x}_{\mathcal{I}}$ durch die sogenannte *Likelihoodfunktion* $P(\mathbf{y}_{\mathcal{I}}|\mathbf{x}_{\mathcal{I}})$ in Beziehung gebracht. Von besonderem Interesse bei der Bestimmung einer optimalen Konfiguration sind Maximum-A-Posteriori-(MAP-)Schätzer, welche durch die Maximierung der Wahrscheinlichkeitsdichte[15]

[14] Für den Fall eines Zufallsfeldes mit diskreten Zuständen.

[15] Es wird angenommen, $P(\mathbf{x}_{\mathcal{I}}|\mathbf{y}_{\mathcal{I}})$ besitzt ein eindeutiges Maximum.

$P(\mathbf{x}_\mathcal{I}|\mathbf{y}_\mathcal{I})$ den Modus mit a-posteriori größter Wahrscheinlichkeit des Zufallsfeldes bestimmen. Unter Ausnutzung der *Regel von Bayes* kann die somit definierte wahrscheinlichste Konfiguration $\mathbf{x}_\mathcal{I}{}^*$ des Zufallsfeldes durch Lösen des Ausdrucks

$$\arg\max_{\mathbf{x}_\mathcal{I}\in\Omega_\mathcal{I}} \{P(\mathbf{y}_\mathcal{I}|\mathbf{x}_\mathcal{I})P(\mathbf{x}_\mathcal{I})\} \tag{2.31}$$

bestimmt werden. In vielen Arbeiten der Bayes'schen Bildanalyse werden hierzu gewisse Modellannahmen getroffen, wovon die der unabhängigen, identisch verteilten Beobachtungen zu den wohl meist genutzten gehört, d. h. $P(\mathbf{y}_\mathcal{I}|\mathbf{x}_\mathcal{I}) = \prod_{n\in\mathcal{I}} P(y_n|\mathbf{x}_n)$. So kann z. B. durch Umformulierung der Likelihoodfunktion

$$P(\mathbf{y}_\mathcal{I}|\mathbf{x}_\mathcal{I}) \propto \exp\left[-\sum_{n\in\mathcal{I}} \mathrm{D}(\mathbf{x}_n)\right] \tag{2.32}$$

mit $\mathrm{D}(\mathbf{x}_n) = \log P(y_n|\mathbf{x}_n)$ und der Modellierung der Prioriverteilung $P(\mathbf{x}_\mathcal{I})$ durch das Potts-Modell aus (2.30), das Schätzproblem aus (2.31) folgendermaßen beschrieben werden:

$$\mathbf{x}_\mathcal{I}{}^* \propto \arg\max \left\{ \exp\left[ -\sum_{\langle n\rangle\in\mathcal{C}_1} \mathrm{D}(\mathbf{x}_n) - \sum_{\langle n,u\rangle\in\mathcal{C}_2} \lambda_{\langle n,u\rangle}\, \delta_{\mathbf{x}_n,\mathbf{x}_u} \right] \right\}. \tag{2.33}$$

Hieraus ist direkt ersichtlich, dass die MAP-Schätzung identisch zu dem Ergebnis der Minimierung des Klammerausdrucks

$$\mathrm{E}(\mathbf{x}_\mathcal{I}) = -\sum_{\langle n\rangle\in\mathcal{C}_1} \mathrm{D}(\mathbf{x}_n) - \sum_{\langle n,u\rangle\in\mathcal{C}_2} \lambda_{\langle n,u\rangle}\, \delta_{\mathbf{x}_n,\mathbf{x}_u} \tag{2.34}$$

im Exponenten von (2.33) ist. Eine kurze Einführung in gängige Optimierungsalgorithmen zur Lösung solcher Energiefunktionale findet sich in Anhang A.1.1. Das in dieser Arbeit verwendete Graphenschnittverfahren[16] zur Bestimmung der Szenenstruktur und Segmentierung auf der Basis einer solchen Darstellung wird im Folgenden näher vorgestellt.

[16] Neben der Graphenschnittoptimierung wurde auch eine Optimierung der genannten Schätzgrößen auf Basis des Belief-Propagation-Verfahrens [Felzenszwalb u. Huttenlocher, 2006; Larsen u. a., 2006] durchgeführt. Auf eine detaillierte Beschreibung des letztgenannten Ansatzes wird im Rahmen dieser Arbeit jedoch verzichtet.

### 2.5.2 Graphenschnittverfahren

Unter Graphenschnitten (engl. *graph-cuts*) versteht man die Überführung der MAP-Schätzung (binärer) Probleme auf das Problem der optimalen Partitionierung eines Graphen. Die Abbildung des oben beschriebenen diskreten Optimierungsproblems in eine graphentheoretische Formulierung erlaubt den Zugriff auf effiziente Lösungsverfahren, die in den letzten Jahren unterschiedlichste Bereiche der Bildverarbeitung maßgeblich beeinflusst haben [Birchfield u. Tomasi, 1999; Kolmogorov u. Zabih, 2001; Raj u. Zabih, 2005; Schoenemann u. Cremers, 2006; Appleton u. Talbot, 2006]. Im Kern geht es dabei um die Übertragung des Energieminimierungsproblems auf das Problem des „*minimalen Schnittes*" aus der Graphentheorie. Ein Graph ist hierbei definiert als ein Tripel $\mathrm{G}(\mathcal{P},\mathcal{E},\mathcal{W})$, wobei $\mathcal{P}$ eine Menge von Knoten und $\mathcal{E}$ eine Menge von Kanten bezeichnet. Durch die Gewichte in $\mathcal{W}$ kann die herkömmliche Definition eines Graphen um weitere Eigenschaften bzgl. der Knoten bzw. der Kanten erweitert werden. In dem Graphen gibt es zusätzlich zwei sogenannte *Abschlussknoten*, die über die Abschlusskanten jeweils mit allen Knoten des restlichen Graphen verknüpft sind und als *Quelle q* und *Senke s* bezeichnet werden. Das Ziel einer erfolgreichen Graphenoptimierung ist die Trennung der Knoten in zwei disjunkte Teilmengen $\mathcal{Q}$ und $\mathcal{S}$ unter den Bedingungen $q \in \mathcal{Q}$, $s \in \mathcal{S}$ und $\mathcal{Q} \cap \mathcal{S} = 0$. In diesem Zusammenhang soll hier das Prinzip minimaler Schnitte in Graphen erläutert und darauf aufbauend effiziente Algorithmen zur Bestimmung solcher Schnitte in Graphen näher vorgestellt werden. Die für eine praktische Anwendung notwendige Erweiterung dieses binären Optimierungsproblems auf mehrere Klassen, d. h. die Partitionierung des Graphen in $J > 2$ Teilmengen, folgt als logische Konsequenz im Anschluss.

#### Wahl der Energiefunktion

Man nehme für den Moment an, bei $\mathbf{l}$ handle es sich um ein binäres Labelfeld, d. h. $\mathcal{L} = \{0,1\}$. Allgemein soll nun das diskrete Optimierungsproblem

$$\hat{\mathbf{l}} = \underset{\mathbf{l}\in\mathcal{L}^N}{\arg\min} \{F(\mathbf{l})\} \tag{2.35}$$

behandelt werden, wobei die allgemeine Funktion $F$ eine Darstellung der Funktionen $F^n$ und $F^{n,u}$ der folgenden Form[17] besitzt:

$$F(\mathbf{l}) = \sum_{n\in\mathcal{P}} F^n(\mathrm{l}_n) + \sum_{n<u} F^{n,u}(\mathrm{l}_n,\mathrm{l}_u)\,. \tag{2.36}$$

[17] Die vorliegende Betrachtung beschränkt sich auf Funktionen, die Element der Funktionsklasse $\mathcal{F}^2$ sind. Allgemeinere Funktionsklassen werden in [Kolmogorov u. Zabih, 2004] näher vorgestellt.

Die Lösung dieses binären Optimierungsproblems auf der Basis von Graphenschnitten ist keinesfalls für jede beliebige Energiefunktion $F$ möglich. In [Boykov u. a., 2001] wird folgende hinreichende Bedingung für die Graphrepräsentierbarkeit von Energiefunktionen auf binären Variablen hergeleitet.

**Theorem 2.1 (Graphrepräsentierbarkeit)**
*Sei $F$ eine Funktion von $N$ binären Variablen. Dann ist $F$ graphrepräsentierbar, falls jeder Term $F^{n,u}$ die folgende Ungleichung erfüllt:*

$$F^{n,u}(0,0) + F^{n,u}(1,1) \leq F^{n,u}(0,1) + F^{n,u}(1,0)\,. \tag{2.37}$$

*Als unmittelbare Folge ist die Summe von zwei graphrepräsentierbaren Funktionen wieder graphrepräsentierbar.*

Die Funktion wird dann auch als regulär bezeichnet. Ein Beweis hierfür ist in [Boykov u. a., 2001] zu finden. Als Beispiel kann das Ising-Modell in (2.29) (mit $\lambda_{\langle n,u\rangle} > 0$) angeführt werden, wofür die einzelnen Summanden regulär sind: $F^{n,u}(0,0) + F^{n,u}(1,1) = -\lambda - \lambda \leq \lambda + \lambda = F^{n,u}(0,1) + F^{n,u}(1,0)$. Handelt es sich um eine reguläre Funktion, kann das Optimierungsproblem effizient durch Graphenschnittverfahren gelöst werden, wie im Folgenden gezeigt wird.

### Minimale Schnitte in Graphen

Hierfür betrachte man den Graphen $\mathrm{G}(\mathcal{P},\mathcal{E},\mathcal{W})$ mit den Knoten $p,r,\ldots \in \mathcal{P}$ und den Kanten $e = \langle p,r\rangle \in \mathcal{E}$. $w(e) \in \mathcal{W}$ drückt das entsprechende, nicht negative Kantengewicht aus. Zusätzlich existieren noch die Abschlussknoten $q \in \mathcal{P}$ und $s \in \mathcal{P}$, die jeweils mit allen anderen Knoten des Graphen verbunden sind. Für Knoten $r$ sei $\mathcal{E}_{\text{in}}(r) = \{\langle p,r\rangle : \langle p,r\rangle \in \mathcal{E}\}$ die Menge der in $r$ einlaufenden Kanten und $\mathcal{E}_{\text{out}}(r) = \{\langle r,p\rangle : \langle r,p\rangle \in \mathcal{E}\}$ die Menge aller von $r$ ausgehenden Kanten.

**Definition 2.7 (Schnitte und Kosten)**
*Ein $\langle q,s\rangle$-Schnitt $c$ eines Graphen $\mathrm{G}(\mathcal{P},\mathcal{E},\mathcal{W})$ ist eine Zerlegung $\{\mathcal{Q},\mathcal{P}\backslash\mathcal{Q}\}$ der Knotenmenge $\mathcal{P}$ mit $\mathcal{Q} \subset \mathcal{P}$, $q \in \mathcal{Q}$ und $s \in \mathcal{P}\backslash\mathcal{Q}$. Die Kosten eines $\langle q,s\rangle$-Schnittes ergeben sich dann aus den „geschnittenen" Kantengewichten $w \in \mathcal{W}$ zu*

$$\mathcal{W}(c) = c(\mathcal{Q},\mathcal{S}) = |c| = \sum_{\substack{p\in\mathcal{Q},r\in\mathcal{S},\\ \langle p,r\rangle\in\mathcal{E}}} w(p,r)\,. \tag{2.38}$$

Ziel eines Graphenschnittverfahrens ist es nun, aus der Menge aller möglichen $\langle q,s\rangle$-Schnitte, den Schnitt mit den geringsten Kosten $|c|_{\min}$ zu bestimmen.

Betrachtet man das oben gestellte Problem als *Flussnetzwerk*, so stellt der Graph $G(\mathcal{P},\mathcal{E},\mathcal{W})$ ein Netzwerk dar. Jede der Kanten $e$ hat ein bestimmte Kapazität. Anschaulich kann man sich eine solche Kante als Rohr vorstellen, die ein bestimmtes Durchleitungsvermögen hat, welches durch die Kapazität $w(e)$ ausgedrückt wird. Das Netzwerk kann somit als Rohrleitungssystem aufgefasst werden. Ausgehend vom Quellknoten $q$ fließt hier Wasser zur Senke $s$. Ein *Fluss* $f$ drückt die Durchleitungsmenge durch das Gesamtnetzwerk aus. Eine Kante gilt als gesättigt, falls $f(e) = w(e)$.

**Definition 2.8 (Flussbedingungen)**
*Folgende Bedingungen werden an den Fluss gestellt, damit $f$ einen gültigen Fluss modelliert:*

*(i) Flusserhaltungsbedingung: Nach dem 1. Kirchhoff'schen Gesetz ist die Menge des einfließenden Wassers gleich der Menge des abfließenden Wassers an einem Knoten, d. h.*

$$\forall\, p \in \mathcal{P}\backslash\{q,s\} : \underbrace{\sum_{e\in\mathcal{E}_{in}(p)} f(e)}_{f^+(p)} = \underbrace{\sum_{e\in\mathcal{E}_{out}(p)} f(e)}_{f^-(p)}. \tag{2.39}$$

*Der Gesamtfluss im Netzwerk ist definiert durch $|f| = f^+(s) - f^-(s)$.*

*(ii) Kapazitätsbedingung: Es kann nie mehr durch eine Kante hindurchfließen als die Kapazität erlaubt, d. h.*

$$\forall\, w(e) \in \mathcal{E} : 0 \le f < w(e). \tag{2.40}$$

Ziel eines Verfahrens zur Bestimmung des maximalen Flusses durch das Netzwerk ist es nun, den maximal möglichen Fluss $|f|_{\max}$ von der Quelle $q$ durch das Flussnetzwerk zur Senke $s$ zu leiten. Dies wird meist schrittweise gemacht, wobei die Idee darin besteht, Pfade von $q$ nach $s$ zu finden, über die es möglich ist, einen höheren Durchfluss zu erreichen. Auf solchen flussvergrößernden Pfaden[18] wird der Fluss erhöht. Das Verfahren endet, wenn kein solcher Pfad mehr existiert. Um die Suche nach flussvergrößernden Pfaden einfach zu gestalten, wird ein *Restflussnetzwerk* eingesetzt.

[18] Im weiteren Verlauf wird hierfür auch der Begriff der flussaugmentierenden Pfade verwendet.

**Definition 2.9 (Restnetzwerk)**
*Zu einem beliebigen Fluss $f$ auf* G *mit der Kapazität $\mathcal{W}$ existiert ein Residualgraph* $G_f(\mathcal{P}, \mathcal{E}_f, \mathcal{W}_f)$*, der die restlichen Kantenkapazitäten bzgl. eines Flusses auf* G *fasst. Der Residualgraph besitzt die gleiche Knotenmenge wie* G *und die Kantenmenge*

$$\mathcal{E}_f = \{e | e \in \mathcal{E}\,,\, f(e) < w(e)\} \cup \{e^{\leftarrow} | e \in \mathcal{E}\,,\, f(e) > 0\}\,. \tag{2.41}$$

Das Restnetzwerk mit Residualgraph $G_f(\mathcal{P}, \mathcal{E}_f, \mathcal{W}_f)$ und Residualkapazitäten $\mathcal{W}_f$ bezüglich eines beliebigen Flusses $f$ zeigt die restlichen Kapazitäten des Netzwerks an. Für jede Kante $e = \langle p, r\rangle \in \mathcal{E}$ mit $f(e) > 0$ enthält $\mathcal{E}_f$ eine Rückkante $e^{\leftarrow} = \langle r, p\rangle$. Die Residualkapazitäten $\mathcal{W}_f$ geben für eine Kante aus $\mathcal{E}$ an, um wie viel der Fluss auf ihr noch erhöht werden kann, also $\forall e \in \mathcal{E} : w_f(e) = w(e) - f(e)$; für eine Rückkante $e^{\leftarrow}$ geben die Residualkapazitäten an, um wie viel der Fluss auf der zugehörigen Hinkante $e$ verringert werden kann, also $\forall e \in \mathcal{E} : w_f(e^{\leftarrow}) = f(e)$.

Eine Möglichkeit um eine Aussage über den maximalen Fluss in Graphen zu treffen bietet die Zerlegung des Graphen in zwei Teile, womit die Verbindung zu den minimalen Schnitten hergestellt werden kann.

**Definition 2.10 (Zerlegung von Flussnetzen)**
*Wenn $f$ ein Fluss in* G *ist, dann bezeichnet*

$$f(\mathcal{Q}, \mathcal{S}) = |f| = \sum_{p \in \mathcal{Q}, r \in \mathcal{S}} f(p, r) \tag{2.42}$$

*den Fluss durch den geschnitten Graphen $(\mathcal{Q}, \mathcal{S})$ und $c(\mathcal{Q}, \mathcal{S})$ seine Kapazität. Obige Gleichung besagt, dass der Wert des Flusses durch den Schnitt genau dem Fluss durch das gesamte Netzwerk entspricht. Weiterhin bildet die Kapazität eines beliebigen Schnitts von* G *eine obere Schranke für den Fluss, d. h. $|f| \leq c(\mathcal{Q}, \mathcal{S})$.*

Die letzte Feststellung ergibt sich aus der Überlegung, dass der Fluss jeder einzelnen, durch den Schnitt laufenden, Kante nach oben beschränkt ist durch die entsprechende Kapazität der jeweiligen Kante. Somit ist auch die Summe beschränkt durch die Kapazität des Schnittes. Die Aussagen über Restnetzwerk, augmentierende Pfade und Schnitte von Graphen können im folgenden Satz in Zusammenhang gebracht werden.

**Theorem 2.2 (Max-Flow/Min-Cut)**
*Sind $\mathcal{Q}$ und $\mathcal{S}$ disjunkte Mengen von Knoten in einem (gerichteten oder ungerichteten) endlichen Netzwerk* G*, ist der maximal mögliche Fluss $|f|_{max}$ von $\mathcal{Q}$ nach $\mathcal{S}$ gleich dem Minimum der Summe der Kapazitäten über alle Schnitte $|c|_{min}$.*

Dann sind folgende Aussagen äquivalent:

- $f$ ist der maximale Fluss in G.
- Das Residualnetzwerk $G_f$ enthält keinen Pfad der den Fluss im Netzwerk erhöhen kann.
- $|f| = c(\mathcal{Q}, \mathcal{S})$ gilt für irgendeinen Schnitt $c$.

Ein Beweis hierfür wird u. a. in [Ford u. Fulkerson, 1956; Mellouli u. Suhl, 2009] gegeben. Der Satz ist das theoretische Fundament für die effiziente Berechnung eines minimalen Graphenschnittes und damit der Szenensegmentierung mit maximal a-posteriori Wahrscheinlichkeit.

### Bestimmung maximaler Flüsse in Graphen

Das Gros der aus der Literatur bekannten *Max-Flow/Min-Cut*-Algorithmen arbeitet entweder nach dem Prinzip der „flusserhöhenden Pfade" nach Ford-Fulkerson [Ford u. Fulkerson, 1956] oder aber des „push-relabeling" nach Goldberg-Tarjan [Goldberg u. Tarjan, 1990].

Beim *Push-Relabeling* besitzt jeder Knoten die Möglichkeit, eine beliebige Menge Flüssigkeit vorübergehend in einem Reservoir zu speichern, d. h. die Flusserhaltungsbedingung wird hier verletzt. Zusätzlich ist zu jedem Knoten die Distanz zur Senke gespeichert. Der Algorithmus versucht nun mittels der „Push"-Operation den Flussüberschuss der entsprechenden Knoten lokal entlang ungesättigter Kanten in Richtung von Knoten mit geringerer Distanz zur Senke abzubauen. Die entsprechende Richtung und Stärke des Flusses wird dabei in Analogie zu dem Modell des Flussnetzwerks durch die relative Höhe des Knotens gesteuert. Mit Ausnahme von Quelle und Senke ist die Höhe aller Knoten variabel: Sammelt sich Flüssigkeit an einem Knoten ohne eine Möglichkeit weiter abzufließen, so wird dieser Knoten angehoben. Die Knotenhöhen bestimmen, wie der Fluss „gepusht" wird, wobei nur Flüsse von höher zu niedriger liegenden Knoten erlaubt sind. Sind Kanten gesättigt, erhöht sich die Distanz des jeweiligen Knotens durch die „Relabel"- Operation. Eine anschauliche Beschreibung des Verfahrens findet sich u. a. in [Cormen u. a., 2001].

Die Mehrheit der Verfahren zur Bestimmung des maximalen Flusses gehören zur Gruppe der *flusserhöhenden Pfadsuchverfahren* auf Basis des Ford-Fulkerson Algorithmus. Hier wird solange nach $\langle q, s \rangle$-Pfaden im Residualgraphen gesucht, die eine Erhöhung des Gesamtflusses zur Folge haben, bis kein solcher flusserhöhender Pfad mehr existiert. Dabei wird die Verteilung des Flusses zwischen den Kanten in dem oben definierten Residualgraph gespeichert. Zu Beginn entsprechen die

Restkapazitäten von $G_f$ den Ausgangskapazitäten von G. Der Fluss $f$ wird mit null initialisiert. Es wird ein Pfad im Residualgraph gesucht, der über ungesättigte Kanten von der Quelle zur Senke führt. Ist ein Pfad gefunden, wird die geringste Kantenkapazität $w$ im Pfad ermittelt. Die Restkapazitäten der Kanten im Pfad werden um die Kantenkapazität $w$ verringert. Die Restkapazitäten der entgegen gerichteten Kanten werden entsprechend um die Kantenkapazität $w$ erhöht. Es wird also der maximal mögliche Fluss durch den Pfad geschickt, der mindestens eine Kante sättigt. Dieser Verarbeitungsschritt wird als Erweiterungsphase (engl. *augmentation stage*) bezeichnet. Jede Erweiterungsphase erhöht den gesamten Fluss um die Kapazität $w$. Dies wird solange durchgeführt, bis der maximale Fluss erreicht ist; also jeder Pfad von der Quelle zur Senke über mindestens eine gesättigte Kante im Residualgraph führt. Zur Optimierung des Rechenaufwands gibt es unterschiedliche Ansätze: Eine Möglichkeit besteht darin, bei der Suche eines Pfades von der Quelle zur Senke, den kürzesten Pfad zu verwenden. Dies kann realisiert werden, indem zu jedem Knoten die Distanz zur Senke gespeichert wird. Ein anderer Ansatz verfolgt die Strategie, die Kapazitäten zu skalieren, da die Komplexität des Verfahrens von der maximalen Kantenkapazität abhängig ist. Ein wesentlicher Nachteil der so arbeitenden Verfahren ist die lange Laufzeit, da jedes Mal eine neue Breitensuche für Pfade von der Quelle zur Senke gestartet wird, bis die Kapazität aller Pfade erschöpft ist. In der Literatur sind diesbezüglich eine Vielzahl von Ansätzen zu finden, die durch den Einsatz verschiedener Heuristiken [Cormen u. a., 2001; Korte u. Vygen, 2008] eine teilweise deutliche Verbesserung der Laufzeit erreichen.

Hervorzuheben ist hier der Max-Flow/Min-Cut Algorithmus von Boykov und Kolmogorov [Boykov u. a., 2001; Boykov u. Kolmogorov, 2001], welcher ebenfalls nach flusserweiternden Pfaden im Graphen von der Quelle zur Senke sucht. Der wesentliche Unterschied zu den oben genannten Verfahren besteht darin, dass es zwei Suchbäume gibt, welche aufeinander zuwachsen. Ein Suchbaum hat seine Wurzel in der Quelle, der andere in der Senke. Die Suchbäume werden wiederverwendet und müssen nicht nach jedem Erweiterungsschritt erneut aufgebaut werden. Im Suchbaum Q mit der Quelle als Wurzel sind alle Kanten von den Elternknoten zu den nachfolgenden Knoten ungesättigt, während im Suchbaum S mit der Senke als Wurzel alle Kanten von den nachfolgenden Knoten zu den Elternknoten ungesättigt sind. Es gibt aktive und passive Knoten. Die aktiven Knoten sind für das Wachsen des Suchbaums zuständig. Ihnen ist es erlaubt freie Nachbarknoten zu akquirieren, die über ungesättigte Kanten verbunden sind. Passive Knoten können nicht neue Knoten akquirieren, da sie von anderen Knoten des gleichen Suchbaums blockiert sind. Ein Pfad ist gefunden, sobald ein aktiver Knoten in einem der Bäume auf einen Nachbarknoten trifft, der zum anderen Baum gehört. In Anhang A.1.2 findet sich hierzu eine detaillierte Beschreibung des Verfahrens. Die

Autoren in [Boykov u. Kolmogorov, 2004] räumen ein, dass die Zeitkomplexität des Verfahrens theoretisch schlechter ist als bei Standardimplementierungen[19] des Max-Flow/Min-Cut-Algorithmus. Jedoch zeigt der experimentelle Vergleich mit den Standardalgorithmen im Anwendungsbereich der Bildverarbeitung eine deutliche Überlegenheit des Verfahrens bzgl. Rechenzeit und Qualität der Ergebnisse.

### 2.5.3 Erweiterung des Verfahrens auf mehrere Klassen

Die Beschränkung auf binäre Probleme galt bei den Graphenschnittverfahren lange Zeit als unüberwindbar, womit eine praktische Nutzung nur auf sehr wenige Anwendungsfälle reduziert blieb. Um diese Einschränkung aufzulösen, werden in [Ishikawa, 2003; Ishikawa u. Geiger, 1999] Bedingungen an das Energiefunktional gestellt, für welche auch bei nicht binären Problemen ein globales Minimum gefunden werden kann. Die Bedingungen schränken die Modellierung jedoch stark ein. So genügt z. B. das in dieser Arbeit verwendete Potts-Modell nicht den dort geforderten Bedingungen.

Ein anderer Ansatz wird in [Ferrari u. a., 1995; Veksler, 1999; Boykov u. a., 2001; Kolmogorov u. Zabih, 2004] verfolgt. Hier wird das Problem minimaler Schnitte in Graphen auf mehr als zwei Abschlussknoten erweitert. Eine solche Formulierung wird im Kontext der Graphenschnitte als „Multiway-Cut" bezeichnet und ist für generelle Graphen NP-schwer [Dahlhaus u. a., 1994], was den Einsatz von Approximationsalgorithmen motiviert. Konkret wird hier die Komplexität einer Mehrklassenoptimierung durch die Kombination mehrerer Klassen zu Superklassen auf ein binäres Optimierungsproblem herunter gebrochen. Ein in der Bildverarbeitung vielverwendetes Verfahren zur approximativen Lösung dieses Problems wird als $\alpha$-Expansion-Algorithmus bezeichnet. Durch eine iterative Optimierungsstrategie, auf welche in Anhang A.1.3 näher eingegangen wird, kann hier eine suboptimale Zerlegung des Graphen in $J > 2$ Klassen erreicht werden. In jedem Iterationsschritt wird ein Teil des Gesamtproblems auf ein binäres Optimierungsproblem abgebildet, welches dann mit dem oben beschriebenen Verfahren gelöst werden kann.

## 2.6 Zusammenfassung

Zur Beschreibung des Abbildungsprozesses der 3D Szene auf die 2D Bildebene wurde das Modell einer idealen Lochkamera vorgestellt. Dieses Modell wurde um eine zweite, räumlich starr dazu versetzte Kamera erweitert, womit prinzipiell eine

[19] Zu nennen sind hier der Dinic- und der Edward-Karps-Algorithmus.

3D Rekonstruktion der statischen Szene möglich wird. Das für die Bewegungsschätzung genutzte parametrische Modell eines sich starr im Raum bewegten Körpers wurde eingeordnet und vorgestellt. Zur Beschreibung unterschiedlich bewegter Objekte in der Szene wurde ein Labelprozess eingeführt, der Gruppen von Bildpunkten zusammenfasst, welche zu einer Objektinstanz gehören. Die stochastische Modellierung räumlicher und zeitlicher Abhängigkeiten der Segmentier- und Rekonstruktionsaufgabe erfolgt mit Hilfe von Markov-Zufallsfeldern. Als grundlegende Forderung für Markov-Zufallsfelder wurde die lokale Charakteristik vorgestellt, mit welcher die Verbundwahrscheinlichkeit einer spezifischen Konfiguration des Zufallsfeldes über die bedingten Wahrscheinlichkeiten der lokalen Entitäten formuliert werden kann. Eine Bestimmung der Verbundwahrscheinlichkeit, auch globale Charakteristik des Zufallsfeldes genannt, aus diesen bedingten Wahrscheinlichkeiten ist jedoch numerisch meist nicht handhabbar. Zur Lösung dieses Problems wurde das Theorem von Hammerley-Clifford eingeführt. Der praktische Wert des Theorems liegt in der Spezifizierung der Verbundwahrscheinlichkeit durch Cliquenpotentiale. Diese Potentialfunktionen bilden die lokale Charakteristik eines Markov'schen Zufallsfeldes auf ein Gibbs-Feld ab, womit a-priori Wissen in den Schätzprozess mit einbracht werden kann. Die Eigenschaft der Homogenität und Isotropie eines Zufallsfeldes erlaubt hier eine effiziente und einfache Berechnung der lokalen Charakteristik bzw. der Cliquenpotentiale, da diese jeweils unabhängig von der Position in $\mathcal{I}$ und Orientierung der Nachbarschaftsbeziehung sind. Abschließend wurde das in dieser Arbeit verwendete globale Optimierungsverfahren der minimalen Graphenschnitte eingeordnet und näher vorgestellt.

KAPITEL 3

# Die Szenensegmentierung

Aufbauend auf den Grundlagen zur Szenensegmentierung wird in diesem Kapitel ein probabilistisch motiviertes Modell zur Auswertung von Stereobildfolgen, hin zu einer bewegungsgestützten Szenensegmentierung, erstellt. Das Ziel der Auswertung ist eine zeitlich konsistente Zerlegung der Bildsequenz in sich nicht überlappende Segmente $\mathcal{X}^1, \mathcal{X}^2, \ldots, \mathcal{X}^J \subset \mathcal{X}$, welche als starr bewegte Objekte im Raum interpretiert werden können. Abbildung 3.1 illustriert dies am Beispiel einer synthetisch generierten Verkehrsszene mit zwei bewegten Objekten relativ zum Kamerasystem. Für jedes der Bildsegmente wird die jeweilige Bewegung durch

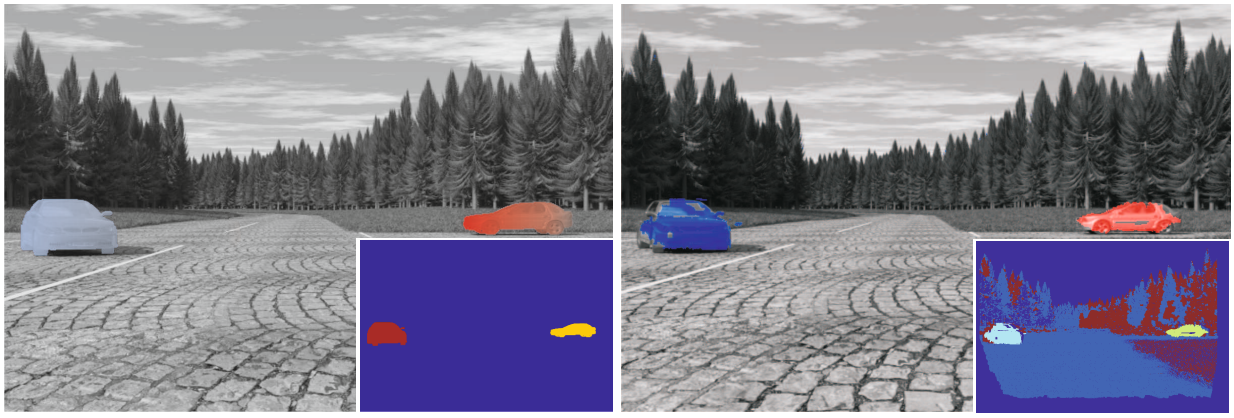


**Abbildung 3.1**: Links: Das Bild zeigt eine synthetisch generierte Verkehrsszene mit zwei fremdbewegten, farblich markierten Objekten. Bildpunkte, die sich mit der Kamera bewegen, sind farblich nicht gekennzeichnet. Unten rechts ist die dazugehörige Segmentierungskarte dargestellt. Rechts: Segmentierergebnis des hier vorgestellten Verfahrens. In dunkelblauen Bereiche im Bild unten rechts ist keine Information für die Segmentierung vorhanden. Dunkelrote Bildpunkte kennzeichnen Bereiche, in denen keine eindeutige Zuweisung möglich ist. Hellblaue Bereiche zeigen relativ zur Kamera statische Bereiche der Szene.

ein Objektmodell beschrieben, welches innerhalb der entsprechenden Bildregion uneingeschränkte Gültigkeit hat und im folgenden Abschnitt genauer spezifiziert wird. Objekte werden dabei als starr bewegte Körper mit bestimmten geometrischen Eigenschaften modelliert. Bezüglich der unmittelbaren Umgebung um das Fahrzeug, wird die Annahme einer ebenen Welt getroffen. Durch die Schätzung der vorausliegenden Fahrbahnebene kann somit eine Vorsegmentierung der Szene durchgeführt werden. Mit dem global im Bild definierten Gütemaß wird für die restlichen Szenenbereiche ein probabilistisches Modell entwickelt, welches die starken statistischen Bindungen der Bildfolge in räumlicher und zeitlicher Richtung explizit in der Form von Glattheitsforderungen an das Ergebnis der Segmentierung und die Rekonstruktion berücksichtigt.

## 3.1 Definition der Segmentieraufgabe

### 3.1.1 Das Objektmodell

Objekte sind in dieser Arbeit definiert als Szenenbereiche, die eine ähnliche Bewegung im Raum aufweisen. Einzelne Bildpunkte der Sequenz, deren Grauwertverschiebung in zeitlicher Richtung ausschließlich mit der Bewegung der Kamera durch eine statische Szene beschrieben werden kann, werden zu einem Objekt zusammengefasst. Im weiteren Verlauf der Arbeit wird für dieses Objekt auch der Begriff *Hintergrundsegment* benutzt. Die Bildpunkte des Hintergrundsegments sind mit dem sogenannten Ego-Label gekennzeichnet. Das entsprechende Bewegungsprofil wird als Eigen- bzw. Egobewegung der Kamera bezeichnet und durch das, in Abschnitt 2.2 vorgestellte, Modell der Starrkörperbewegung beschrieben. Alle weiteren Bewegungsmuster in der Szene werden relativ zu dieser Eigenbewegung untersucht. Bereiche im Bild, die sich relativ zum Kamerasystem fremd bewegen, werden auf ihre Ähnlichkeit hinsichtlich einer starren Bewegung im Raum untersucht und ggf. einer neuen Objekthypothese zugewiesen. Unabhängig für jede neue Objekthypothese wird die Bewegung aus den Bilddaten geschätzt und in den Segmentierprozess zurückgeführt. Als Schätzgröße ergibt sich hieraus nach (2.9) für jede Hypothese der Parametervektor $\mathbf{v}_t = (\omega_x, \omega_y, \omega_z, t_x, t_y, t_z)$. Mit dem Ziel einer vollständig parametrischen Beschreibung der Szenenobjekte werden für die *fremdbewegten* Objektinstanzen zusätzliche geometrische Annahmen eingeführt. Die Objektgeometrie sei im Weiteren ausgedrückt durch den Parametervektor $\boldsymbol{\xi}_t$, wobei die Beschreibung beliebige Modelle beinhaltet. Denkbar ist hier u. a. die Modellierung der Objektgeometrie durch Ebenen [Adiv, 1985; Li u. Zucker, 2005], Polygone [Schindler, 2005] oder auch kompliziertere Freiformflächen. Als einfaches Modell zur Beschreibung der

Geometrie, wird in dieser Arbeit jedes Objektsegment durch seine relative Lage $\mathbf{M}_t = (X, Y, Z)$ im 3D kartesischen Raum mit einer entsprechenden Ausdehnung $\mathbf{\Sigma}_t$ definiert, d. h. $\boldsymbol{\xi}_t = (\mathbf{M}_t, \mathbf{\Sigma}_t)$. Beide Größen werden aus der Segmentierung bestimmt. Durch die Berücksichtigung der Objektgeometrie ergibt sich ein vollständig parametrisches Modell mit Parametern $\boldsymbol{\theta}_t = (\mathbf{v}_t, \boldsymbol{\xi}_t)^{\mathrm{T}}$, welches den entsprechend gültigen Bildbereich effizient durch wenige Parameter näherungsweise beschreibt. Mit einem solches Modell kann auch in schwach bzw. nicht texturierten Bildbereichen eine gültige Bewegungsschätzung erfolgen, vorausgesetzt das Bildsegment enthält zumindest einige signifikante Texturmuster.

### 3.1.2 Geometrische Modellierung der Szene

Bezüglich der statischen Szene relativ zum Beobachter erscheint, aus Gründen der darin enthaltenen komplexen Strukturen, eine vollständig parametrische Beschreibung als impraktikabel. Einzige Ausnahme stellt hier die vorausliegende Fahrbahn dar. Laut dem „*wahrnehmungsökologischen Erklärungsansatz*" von Gibson [Gibson, 1950], stellt die Wahrnehmung des (ebenen) Untergrunds (engl. *ground plane perception*) der Szene eine der elementaren Voraussetzungen der menschlichen Mobilität dar. Nach seiner Bodentheorie (engl. *ground theory*) ist die wesentliche Information für die visuelle Wahrnehmung räumlicher Tiefe und Entfernung nicht durch die Objekte selbst, sondern durch die Anordnung des Untergrunds gegeben („*there is literally no such thing as a perception of space without the perception of a continuous background surface*"). Im Bereich der mobilen Umfeldwahrnehmung wird diese intuitive Feststellung vielfach genutzt, um relevante Bereiche in der vorausliegenden Fahrzeugumgebung zu bestimmen und entsprechend Aktionen daraus abzuleiten [Se u. Brady, 2002; Leibe u. a., 2007; Ess u. a., 2009]. Ein einfaches, aber meist ausreichendes geometrisches Modell zur Beschreibung der unmittelbaren Fahrzeugumgebung, ist dabei die Ebene. Als relevante Bereiche gelten in diesem Zusammenhang meist die Teile einer Szene, die aus der geschätzten Fläche der vorausliegenden Fahrzeugumgebung herausragen bzw. darauf platziert sind. Wird ein kalibriertes Stereokamerasytem verwendet, kommt häufig die sog. *v-Disparität* [Labayrade u. a., 2002] zum Einsatz. Hierbei wird angenommen, dass sich die Szene aus einer Menge von Ebenen zusammensetzt, die horizontal oder vertikal im Raum orientiert sind [Soquet u. a., 2007]. Ist diese Annahme erfüllt, werden die einzelnen Ebenen der realen Szene im v-Disparitätsbild auf Linien abgebildet. Vertikale Ebenen im Bild werden hierbei durch vertikale Linien, horizontale Ebenen entsprechend durch geneigte Linien repräsentiert. Die Bestimmung der v-Disparität aus dem Disparitätsbild erfolgt durch zeilenweise Akkumulation der Disparitätswerte. Da die Fahrbahnebene als dominante horizontale Ebene im Bild vorausgesetzt wird, kann aus den aufakkumulierten Werten

die sog. *ground correlation line* [Broggi u. a., 2005] eindeutig identifiziert und parametriert werden. Hindernisse, die sich auf der Fahrbahn befinden, können dann durch einfache Verfahren detektiert werden. Da die v-Disparität jedoch sehr anfällig auf Wank- und Gierwinkeländerungen der bewegten Kamera ist, wurde alternativ dazu ein weiteres Verfahren zur Schätzung der Fahrbahngeometrie aus Stereobilddaten entwickelt. Zur Beschreibung der Fahrbahn können hier grundsätzlich beliebige Geometriemodelle verwendet werden, wobei sich gezeigt hat, dass in den meisten praktischen Fällen das Ebenenmodell eine hinreichende Näherung der tatsächlichen Oberflächengeometrie darstellt. Durch die Transformation der Ebene in den Bildbereich, kann bei dem in Anhang A.2 näher beschriebenen Verfahren die hohe Fehlerempfindlichkeit bei der Ebenenschätzung mit zunehmender Szenentiefe verkleinert werden. Für die Schätzung wurde ein Total-Least-Squares-Ansatz gewählt, der die orthogonalen Abstände zwischen gemessenem Ebenenpunkt und korrespondierendem Modellpunkt minimiert. Um die Schätzung robust gegen Ausreißer[1] zu machen, wurde ein Least-Median-of-Squares-Verfahren in Kombination mit einem M-Estimator[2] gewählt, welches bis zu einem Anteil von 50% an Ausreißern gute Ergebnisse liefert. Die Bewertung des Verfahrens hinsichtlich Schätzgüte und Robustheit erfolgt in Kapitel 5.

Mit Hilfe der geschätzten Fahrbahnebene, kann für jeden Szenenpunkt dessen Höhe über Grund bestimmt werden, wie in Abbildung 3.2 dargestellt. Punkte, die unterhalb bzw. auf der Fahrbahnebene liegen, werden in der weiteren Betrachtung als statisch klassifiziert, da erwartet werden kann, dass sich in diesen Szenenbereichen keine fremdbewegten Objekte befinden. Punkte, die oberhalb einer bestimmten Höhe[3] zur geschätzten Fahrbahnebene liegen, werden bei der Bewegungssegmentierung ebenfalls als statisch angenommen. Die Ebeneninformation wird außerdem zur Bestimmung des relativen Nickwinkels $\nu$ und der relativen Höhe $h$ der Kamera über der geschätzten Fahrbahn genutzt.

In Abbildung 3.3 sind die einzelnen Verarbeitungsschritte des Verfahrens und das vorgestellte Szenenmodell veranschaulicht dargestellt. Beim Ergebnis der Segmentierung (IV) ist erkennbar, dass die beiden Fahrzeuge links im Bild zu einem Objekt zusammengefasst werden. Aufgrund der ähnlichen Bewegung der beiden Fahrzeuge und ihrer räumlichen Nähe zueinander ist eine Trennung zu diesem Zeitpunkt noch nicht möglich.

[1] Ausreißer sind im Zusammenhang mit der Schätzung der Fahrbahnebene definiert als Bereiche, die auf der zu schätzenden Fahrbahn liegen.

[2] Von Maximum-Likelihood-*artig*. Diese Klasse von Schätzfunktion kann als Verallgemeinerung der Maximum-Likelihood-Methode angesehen werden [Huber, 2004].

[3] Nach §32 Abs. 2 der Straßenverkehrs-Zulassungs-Ordnung (StVZO) liegt die maximal zulässige Höhe eines Kraftfahrzeuges einschließlich Anhänger bei $4,00$m.

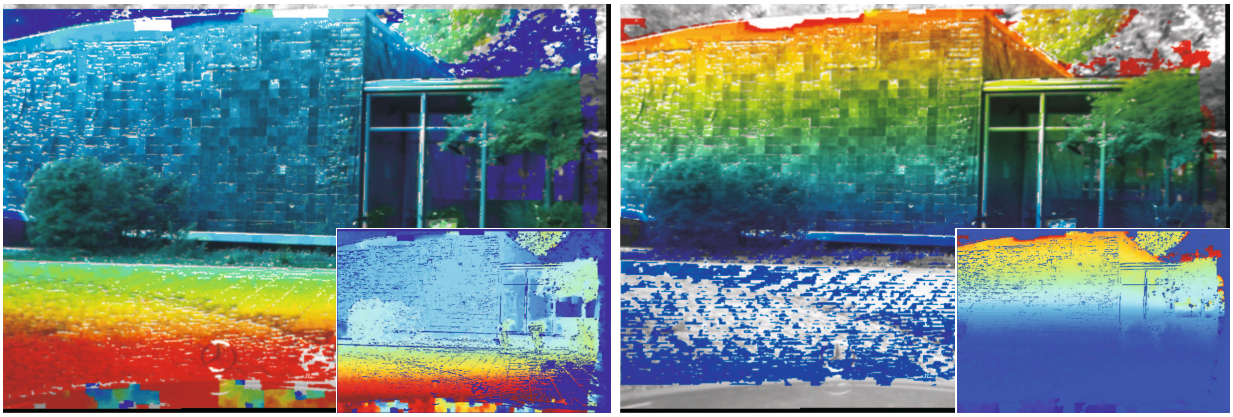


**Abbildung 3.2**: Links: Ausgabe der Stereorekonstruktion. Die Tiefe der Szene, d. h. der relative Abstand zum Kamerasystem, ist farblich codiert: Rot symbolisiert Nähe; mit zunehmender Tiefe geht die Farbe immer mehr ins Blaue über. In dunkelblauen Bereichen konnte kein Tiefenwert zugeordnet werden. Rechts: Die Höhenkarte der Szene. Sie enthält für jeden Punkt die metrische Höhe relativ zur geschätzten Fahrbahnebene. In der Karte nehmen die Werte, vom Blauen ausgehend, linear zu und gehen mit anwachsender Höhe ins Rote über. Wie bei der Tiefenkarte, ist dunkelblauen Bildpunkten keine Höhe zugeordnet. Punkte mit einer Höhe $h \leq 0\,\mathrm{m}$ sind ausgeblendet.

### 3.1.3 Gütemaß der Segmentierung

Geht man für den Augenblick von einer optimalen Schätzung der Modellparameter aus, kann auf der Basis des resultierenden Verschiebungsfeldes $\mathcal{D}_{\tau\in\mathcal{T}}$ durch *Bewegungskompensation* theoretisch ein *Prädiktionsbild* $\hat{\mathrm{g}}_t(\mathcal{X}) = \mathrm{g}_\tau(\mathcal{X}+\mathcal{D}_\tau)$ aus jedem vorhergehenden[4] Bild $\mathrm{g}_{\tau<t}(\mathcal{X})$ einer orts- und zeitdiskreten Bildfolge $\mathcal{G}_T$ erstellt werden. Für ein optimal geschätztes Verschiebungsfeld ergibt sich daraus, unter Annahme konstanter Grauwerte entlang der Bewegungstrajektorien, ein Prädiktionsbild, welches identisch ist mit dem gemessenen Bild[5] $\mathrm{g}_t(\mathcal{X})$ zum Zeitpunkt $t$. Unter idealen Bedingungen und einer optimaler Schätzung des Bewegungsfeldes, ergibt sich somit ein Prädiktionsfehler von null für jede Stelle im Bild. In der Praxis weicht die bewegungskompensierte Prädiktion jedoch aufgrund von verschiedenen Einflüssen von den gemessenen Grauwerten ab, womit auch der Fehler maßgeblich beeinflusst wird. Eine Quelle für mögliche Abweichungen stellt

[4] Das Gleiche gilt auch in umgekehrter Richtung für die Rückwärtsprädiktion des Bildes, d. h. $t < \tau$, jedoch soll bei der weiteren Betrachtung darauf verzichtet werden.

[5] Für Bilddaten wird im Weiteren außerdem die abkürzende Schreibweise $\mathrm{g}_t(\mathcal{X}) = \mathrm{g}_t$ verwendet. Entsprechend wird auch für andere orts- und zeitdiskrete Größen auf eine solche kompakte Darstellung zurückgegriffen. Es sei weiterhin in Erinnerung gerufen, dass bei Größen, die sich auf die rechte Kamera beziehen, auf den Index r verzichtet wird.

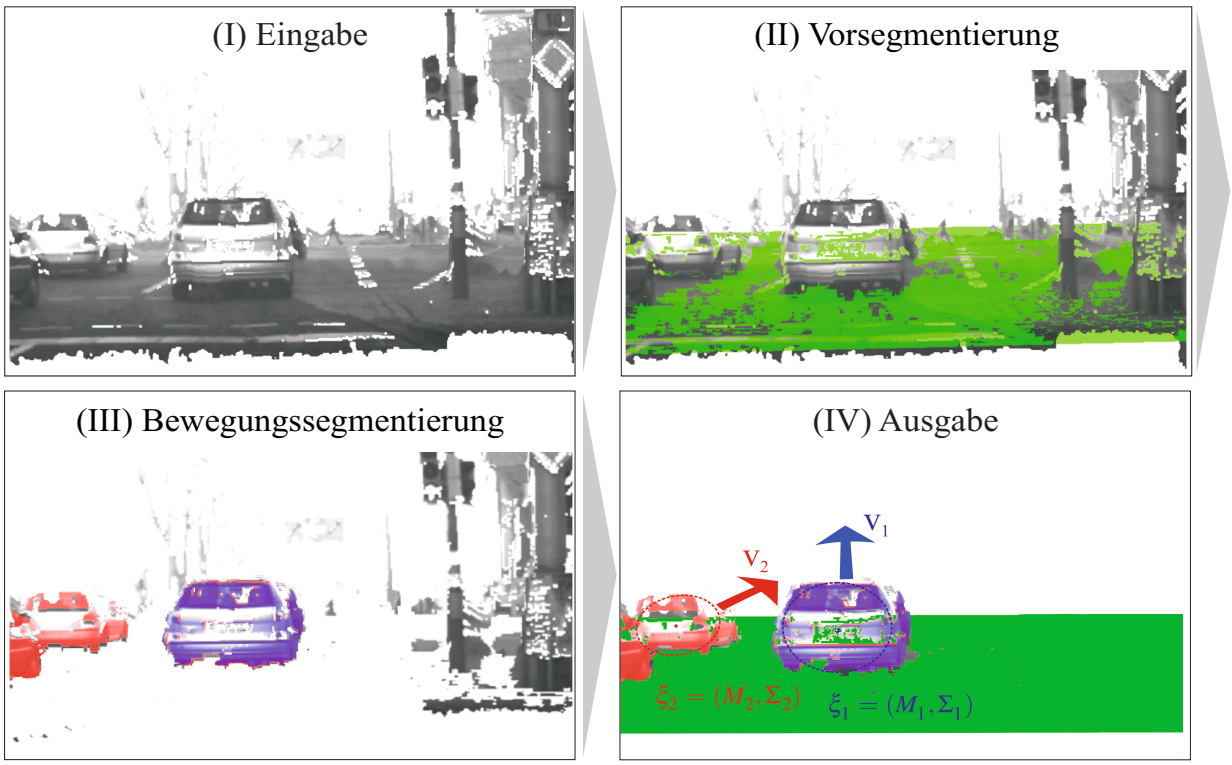



**Abbildung 3.3**: (I) Typische Verkehrsszene mit statischen und fremdbewegten Objekten. (II) Nach Kompensation der Szenenpunkte, die der Fahrbahn zugewiesen werden können, verbleiben lediglich Punkte mit eine Höhe $h > 0$ bzgl. der geschätzten Fahrbahnebene. (III) Aufbauend auf die Vorsegmentierung erfolgt schließlich die Bewegungssegmentierung basierend auf dem Objektmodell aus Abschnitt 3.1.1. (IV) Das Ergebnis der Szenensegmentierung. Hierbei sind Punkte mit einer relativen Tiefe $z_t > 50$m ebenfalls aus dem Bild gelöscht.

die fehlerhafte bzw. unsicherheitsbehaftete Schätzung des Bewegungsfeldes selbst dar. Andere Einflussquellen sind z. B. das Bildrauschen oder Modellfehler, die das Ergebnis systematisch verfälschen. So ist der Fehler in verdeckt werdenden Bereichen zeitlich und räumlich getrennt aufgenommener Bilder nicht sinnvoll definiert. Außerdem wird in natürlichen Szenen die Annahme der Grauwertkonsistenz oftmals verletzt. Im weiteren Verlauf dieser Arbeit werden Modelle entwickelt, welche den Einfluss dieser Fehler auf ein Minimum reduzieren und eine Verwendung der Bewegungskompensation zulassen. Den Einflüssen durch Mess- und Schätzunsicherheiten bei der bewegungskompensierten Prädiktion $\hat{\mathrm{g}}_{t+1} = \hat{\mathrm{g}}_{t+1}(\mathcal{D}_t, \mathrm{g}_t)$ wird durch die Überlagerung additiven Rauschens Rechnung getragen. Dies geschieht in Form einer *Restfehlerkarte* $\mathbf{E}_t^{\mathrm{b}}$, womit für das Folgebild $\mathrm{g}_{t+1}$ ein stochastisches Modell definiert ist, welches in Abschnitt 3.3 näher vorgestellt wird.

Neben der Bewegungsinformation selbst, soll das Gütemaß für die Segmentierung auch die Qualität der Szenenrekonstruktion bewerten. Hier können obige Überle-

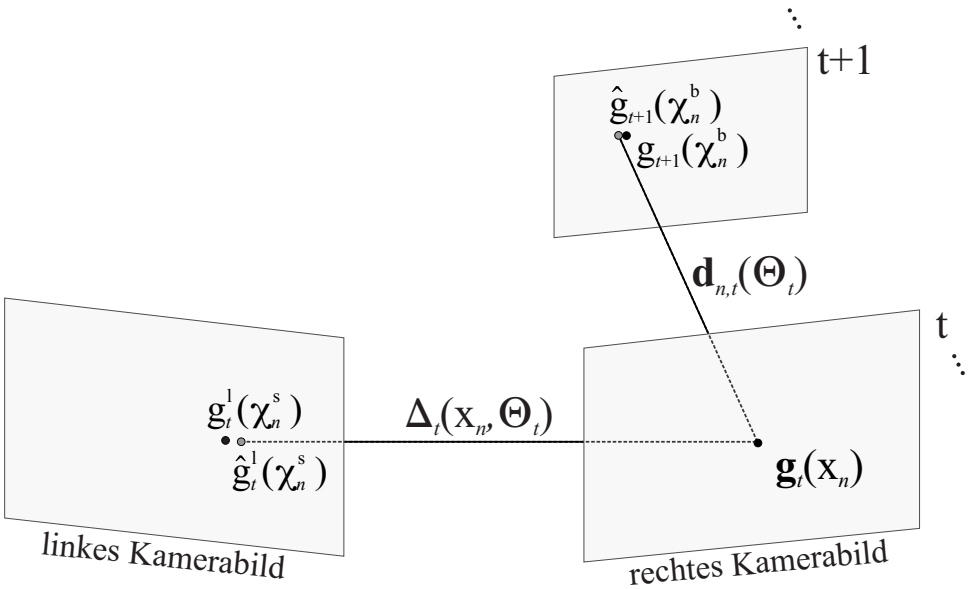



**Abbildung 3.4**: Zum Zeitpunkt $t$ werden die durch den Parametersatz $\boldsymbol{\Theta}_t$ spezifizierten Grauwertverschiebungen in örtlicher und zeitlicher Richtung ausgewertet. Der Fehler geht gegen Null für den Fall, dass innerhalb jedes Bildsegments die Bildbewegung $\mathbf{d}_{n,t}(\boldsymbol{\Theta}_t) \in \mathcal{D}_t$ und die Szenenstruktur $z_{n,t} \in \mathbf{z}_t$ für jeden Punkt $\mathbf{x}_{n\in\mathcal{R}}$ richtig geschätzt wurden.

gungen der bewegungskompensierten Prädiktion ohne Weiteres auf die Bewertung der Szenenrekonstruktion übertragen werden. Wie in Abschnitt 2.2 gezeigt wurde, kann die Szenenstruktur $\mathbf{Z}_t$ mit Hilfe von (2.17) sehr einfach anhand einer Disparitätskarte berechnet werden. Ähnlich zu den Ausführungen bei der Bewegungskompensation kann nun für eine bestimmte Konfiguration $\mathbf{Z}_t = \mathbf{z}_t$ ein Prädiktionsbild $\hat{\mathrm{g}}_t^{\mathrm{l}} = \hat{\mathrm{g}}_t^{\mathrm{l}}(\mathbf{z}_t, \mathrm{g}_t)$ erstellt werden, welches die Grauwerte der rechten Kamera auf die linke Kamera abbildet. Rauscheinflüsse, etc. werden auch hier durch die Addition einer Restfehlerkarte $\mathbf{E}_t^{\mathrm{s}}$ modelliert. Diese Erweiterung macht vor allem dann Sinn, wenn bei der Bestimmung der Szenenstruktur auch die zeitliche Entwicklung der Tiefenwerte mit berücksichtigt wird, wie in Abschnitt 3.4 beschrieben. Bei der Bewegungskompensation sowie bei der Szenenrekonstruktion ist ein Restfehler der nahe bei null liegt wünschenswert. Zur Verdeutlichung sind in Abbildung 3.4 die einzelnen Größen nochmals grafisch dargestellt. Es wird deutlich, dass die Segmentierung der Szene systembedingt immer einen Zeitschritt in der Vergangenheit erfolgt. Aufgrund der hohen Bildrate ist dieser zeitliche Versatz jedoch akzeptabel. Für jede Objekthypothese wird angenommen, dass die jeweilige Bildregion hinreichend genau durch den Parametervektor $\boldsymbol{\theta}_t^j$ charakterisiert ist. Die Gesamtheit aller initialisierten Parametervektoren wird zusammenfassend durch $\boldsymbol{\theta}_t = (\boldsymbol{\theta}_t^1, \ldots, \boldsymbol{\theta}_t^J)$ ausgedrückt. Mit obigem Modell wird jedes bewegte Objekt in der Szene als stationär relativ zur bewegten Kamera beschrieben, d. h. die

gesamte Szenenbewegung wird durch eine Menge von sich starr bewegenden Objekten ausgedrückt, die sich relativ zum kamerafesten Koordinatensystem bewegen. Um eine dichte 3D Szenensegmentierung zu erreichen, wird zusätzlich zu den Bewegungs- und Lageparametern die Szenenstruktur $\mathbf{Z}_t$ in das Modell mit aufgenommen. Eine Szene wird somit durch den Parametersatz $\boldsymbol{\Theta}_t = \{\boldsymbol{\theta}_t, \mathbf{Z}_t\}$ vollständig beschrieben. Für die Segmentierung ergibt sich hieraus folgende Definition der Restfehlerkarte

$$\mathbf{E}_t^2(\mathcal{X}) = \underbrace{\left(\mathrm{g}_{t+1}(\mathcal{X}) - \hat{\mathrm{g}}_{t+1}(\mathcal{X}, \boldsymbol{\Theta}_t, \mathrm{g}_t)\right)^2}_{\mathbf{E}_t^{\mathrm{b}}} + \underbrace{\left(\mathrm{g}_t^{\mathrm{I}}(\mathcal{X}) - \hat{\mathrm{g}}_t^{\mathrm{I}}(\mathcal{X}, \boldsymbol{\Theta}_t, \mathrm{g}_t)\right)^2}_{\mathbf{E}_t^{\mathrm{s}}}. \tag{3.1}$$

Bei der Herleitung des probabilistischen Modells der Szenensegmentierung im weiteren Verlauf dieses Kapitels soll $\boldsymbol{\Theta}_t$ zunächst als bekannt angenommen werden. Durch die Auswertung der Restfehlerkarte kann dann eine *vollständig überwachte* (engl. *fully supervised*) Segmentierung [Li, 2009] der Szene durchgeführt werden. Diese Beschreibung wird in eine iterative Optimierungsstrategie eingebettet, die in alternierender Reihenfolge zuerst die unbekannten Modellparameter $\boldsymbol{\Theta}_t$ schätzt und – darauf aufbauend – die Segmentierung $\mathbf{L}_t$ daraus ableitet. Durch die Auswertung des aktuellen Segmentierergebnisses, wird zusätzlich dazu noch die Anzahl $J_t$ der sich gegenwärtig in der Szene befindlichen, unabhängig bewegten Objekte angepasst. Per Definition handelt es sich dann um ein *vollständig unüberwachtes* Segmentierverfahren. Als zusätzliche Schätzgrößen der Szenensegmentierung ergeben sich hieraus:

(i) Für jede Objekthypothese $j \in \mathcal{J}$ unabhängig, die Bewegung und Lage des Objekts im 3D Raum relativ zum kamerafesten Koordinatensystem. Dies wird durch den Parametersatz $\boldsymbol{\theta}_t^j \subset \boldsymbol{\Theta}_t$ ausgedrückt.

(ii) Die Szenenstruktur $\mathbf{Z}_t$ für möglichst alle im Bild sichtbaren Szenenpunkte. Hierfür steht ein vollständig kalibriertes Stereokamerasystems zur Verfügung.

(iii) Die Anzahl $J_t$ der sich aktuell in der Szene befindlichen, unabhängig bewegten Objekte.

## 3.2 Modellbasierte Bayes-Formulierung

Bei Kenntnis von $\boldsymbol{\Theta}_t$ und $J_t$ ist (3.1) vollständig definiert und es ist möglich, daraus eine Segmentierung $\mathbf{L}(\mathcal{X}, t)$, kurz $\mathbf{L}_t$, abzuleiten. Jedes Bildsegment wird dabei durch eine spezielle Verteilungsfunktion modelliert, welche durch die oben

vorgestellten Modellparameter charakterisiert ist. Diese Segmentierung unterteilt das Bildraster $\mathcal{X}$ in zusammenhängende, sich nicht überlappende Segmente. Dies geschieht in Form eines Labelprozesses $\mathbf{l}(\mathcal{X},t)$, kurz $\mathbf{l}_t$, welcher jeden Bildpunkt $\mathbf{x}_{n\in\mathcal{R}}$ eindeutig einem Segment $\mathcal{X}^{\mathrm{l}_{n,t}\in\mathbf{l}_t}$ zuordnet. Die Glattheit und Kontinuität der Segmentierung $\mathbf{L}_t$ in räumlicher wie zeitlicher Richtung wird durch die stochastische Modellierung als Gibbs/Markov-Zufallsfeld erreicht.

### 3.2.1 Formale Beschreibung der Schätzaufgabe

Für die optimale Schätzung $\mathbf{l}_t^* \in \mathcal{F}$ des wahren Labelfeldes $\mathbf{L}_t = \mathbf{l}_t \in \mathcal{F}$ auf Basis einer Stereobildfolge $\mathcal{G}_T$ der Länge $T = t+1$, wird eine Kostenfunktion $\mathrm{C_l}(\mathbf{l}_t, \hat{\mathbf{l}}_t)$ aufgestellt, deren Erwartungswert durch die Schätzung gerade minimiert wird [Marroquin u. a., 1987]. Als Maß für die Güte der Segmentierung $\hat{\mathbf{l}}_t = (l_1,\ldots,l_n,\ldots,l_N)$, wird in dieser Arbeit die Anzahl der falsch zugewiesenen Elemente genutzt. Durch die harte Zuordnung der jeweiligen Labelwerte im Bayes'schen Sinne, wird jede Fehlentscheidung bzgl. eines falsch zugewiesenen Elements mit gleichen Kosten bestraft. Nach [Schuermann u. Kreßel, 1992; Li, 2009] kann daraus dann die Konfiguration mit minimalen Kosten bestimmt werden, indem man das Labelfeld mit der a-posteriori maximalen Wahrscheinlichkeit schätzt. Unter Berücksichtigung des oben vorgestellten Szenenmodells, wählt der Schätzer somit die Realisation des Labelfeldes, welche bzgl. der Wahrscheinlichkeitsverteilung[6]

$$P(\mathbf{l}_t, \boldsymbol{\Theta}_t | \mathcal{G}_{t+1}) = P(\mathbf{l}_t, \boldsymbol{\Theta}_t | \mathcal{G}^{\mathrm{r}}_{t+1}, \mathcal{G}^{\mathrm{l}}_t) \tag{3.2}$$

ein Maximum aufweist. $\mathcal{G}^{\mathrm{r}}_{t+1}$ und $\mathcal{G}^{\mathrm{l}}_t$ beschreiben hierbei jeweils die Grauwertinformation des rechten und linken Kamerabildes der Stereosequenz. Mit Hilfe des *Satzes von Bayes* kann die a-posteriori Verteilung von (3.2) formal beschrieben werden mit

$$P(\mathbf{l}_t, \boldsymbol{\Theta}_t | \mathcal{G}_{t+1}) = \frac{P(\mathrm{g}_{t+1}, \mathrm{g}^{\mathrm{l}}_t | \mathbf{l}_t, \boldsymbol{\Theta}_t, \mathcal{G}_A) P(\mathbf{l}_t, \boldsymbol{\Theta}_t | \mathcal{G}_A)}{P(\mathrm{g}_{t+1}, \mathrm{g}^{\mathrm{l}}_t | \mathcal{G}_A)}\,, \tag{3.3}$$

wobei aus Gründen der besseren Lesbarkeit $\mathcal{G}_A = \{\mathcal{G}^{\mathrm{r}}_t, \mathcal{G}^{\mathrm{l}}_{t-1}\}$ gewählt wird. Für die weitere Betrachtung wird angenommen, dass die zeitlich zurückliegenden und als konstant anzunehmenden Schätzgrößen $\hat{\mathbf{l}}_{t-1}$ und $\hat{\boldsymbol{\Theta}}_{t-1}$ in dieser Größe implizit enthalten sind.

[6] Wie in Abschnitt 3.1.3 bereits erläutert, liegt die aktuelle Schätzung der Szenensegmentierung zum Zeitpunkt $t$ immer einen Zeitschritt in der Vergangenheit zurück. Die Information des entsprechenden linken Bildes $\mathrm{g}^{\mathrm{l}}_{t+1}$ wird bei der weiteren Betrachtung vernachlässigt.

Der erste Faktor in (3.3) wird als *Likelihood*[7] bezeichnet und bewertet die Beobachtungen für eine gegebene Verteilung der Schätzgrößen. Eine quantitative Formulierung der Likelihood erfolgt in Abschnitt 3.3. Der zweite Faktor enthält Information über die Schätzgrößen selbst, die nicht von den Beobachtungen abhängen und a-priori bekannt sind. Auf die hier entwickelte Modellierung von Vorwissen in Form von Glattheitsanforderungen in räumlicher und zeitlicher Richtung wird in Abschnitt 3.4 näher eingegangen. Durch weitere Umformung des Ausdrucks ergibt sich

$$P(\mathbf{l}_t, \boldsymbol{\Theta}_t | \mathcal{G}_A) = P(\boldsymbol{\Theta}_t | \mathcal{G}_A) P(\mathbf{l}_t | \boldsymbol{\Theta}_t, \mathcal{G}_A) = P(\mathbf{l}_t | \mathcal{G}_A) P(\boldsymbol{\Theta}_t | \mathbf{l}_t, \mathcal{G}_A) \,. \tag{3.4}$$

Mit den oben getroffenen Annahmen ergibt sich aus der Tatsache, dass der Nenner in (3.3) als konstant bzgl. der Labelvariable $\mathbf{l}_t$ ist, dann folgende Formulierung für die Segmentierung

$$\hat{\mathbf{l}}_{t,\mathrm{MAP}} = \arg\max_{\mathbf{l}_t} \left\{ P(\mathrm{g}_{t+1}, \mathrm{g}_t^{\mathrm{l}} | \mathbf{l}_t, \boldsymbol{\Theta}_t, \mathcal{G}_A) P(\mathbf{l}_t | \boldsymbol{\Theta}_t, \mathcal{G}_A) \right\} . \tag{3.5}$$

Auf die Lösung von (3.5) mit Hilfe aktueller Optimierungsverfahren wird in Abschnitt 3.5 eingegangen. Zur vollständigen Beschreibung der Segmentieraufgabe, muss der Ausdruck in (3.2) jedoch auch bzgl. der in der Regel unbekannten Größe $\boldsymbol{\Theta}_t$ optimiert werden. Eine zeitgleiche Bestimmung der Zufallsgrößen $\mathbf{l}_t$ und $\boldsymbol{\Theta}_t$ ist in analytisch geschlossener Form jedoch nicht ohne Weiteres möglich, da die zu schätzenden Größen sich gegenseitig beeinflussen. Mit Hinblick auf eine effiziente Umsetzbarkeit der Szenensegmentierung, wurde in dieser Arbeit eine Strategie entwickelt, welche die verkoppelten Optimierung mehrerer gegenseitig bedingender Zufallsgrößen als Schätzaufgabe mit unvollständigen Beobachtungen (engl. *missing data problem*) beschreibt [Dempster u. a., 1977; Gauvrit u. a., 1997].

### 3.2.2 Schätzung mit unvollständigen Beobachtungen

Die Beobachtungen werden hierbei als *unvollständig* in Bezug auf die Schätzung der Parameter des Szenenmodells bezeichnet. Die Unvollständigkeit der Daten ergibt sich aus der Tatsache, dass sich ein Bild aus einer unbekannten Menge unterschiedlich bewegter Objekte zusammensetzt und die Zugehörigkeit der einzelnen Bildpunkte zu diesen Objekten unbekannt ist.

[7] Im deutschen Sprachgebrauch hat sich der Begriff Likelihood im Bereich der statistischen Signalauswertung mittlerweile etabliert, wobei die Begriffe *Plausibilität* oder *Mutmaßlichkeit* der englischen Entsprechung am nächsten kommen.

Der Labelprozess $\mathbf{l}_t$ repräsentiert hierbei die fehlenden Daten bezüglich der Beobachtungen. Eine iterative Lösungsstrategie auf der Basis des Expectation-Maximization-(EM-)Verfahrens von [Dempster u. a., 1977] wird in Kapitel 4 ausführlich beschrieben. Der Lageparameter $\boldsymbol{\xi}_t \subset \boldsymbol{\theta}_t$ jeder Objekthypothese wird dabei durch eine gewichtete Maximum-Likelihood Schätzung bestimmt. Für die jeweiligen Bewegungsparameter $\mathbf{v}_t \subset \boldsymbol{\theta}_t$ wird ein rekursiver Schätzer beschrieben, der die zeitlich veränderlichen Objektparameter schritthaltend nachführt. Die für eine robuste Schätzung der einzelnen Hypothesenparameter notwendige Datenzuweisung, erfolgt durch die probabilistische Gewichtung aller Beobachtungen [Bachmann, 2009; Bachmann u. Kuehne, 2009]. Hierbei stellt das Labelfeld $\mathbf{l}_t$ eine Art Assoziationsvariable zwischen Daten- und Hypothesenraum dar.

Die Abhängigkeit der Segmentierung von $\mathbf{Z}_t \subset \boldsymbol{\Theta}_t$ erfordert weiterhin die Schätzung der Szenenstruktur. Aufgrund der formalen Ähnlichkeit der Modelle zur Beschreibung einer dichten Szenenrekonstruktion und Segmentierung ist es naheliegend, die beiden Modelle gemeinsam herzuleiten. Aus diesem Grund wird im Rahmen dieses Kapitels, neben der Modellierung der Segmentieraufgabe, auch auf das Modell der Szenenrekonstruktion eingegangen. Der hierbei zu optimierende Ausdruck ergibt sich mit (3.3) und (3.4) zu

$$\hat{\mathbf{z}}_{t,\mathrm{MAP}} = \arg\max_{\mathbf{z}_t} \left\{ P(\mathrm{g}_{t+1}, \mathrm{g}_t^1 | \mathbf{z}_t, \mathbf{l}_t, \mathcal{G}_A) P(\mathbf{z}_t | \mathbf{l}_t, \mathcal{G}_A) \right\} . \tag{3.6}$$

Da sich die Berechnung des Gütemaßes dabei über den gesamten Zustandsraum der Zufallsvariable erstreckt, können auch globale Bedingungen mit in den Schätzprozess eingebunden werden. Diese Bedingungen helfen in Bereichen, in denen die Beobachtungen selbst keine eindeutige Zuweisung erlauben, durch Modellannahmen eine plausible Szenenoberfläche zu rekonstruieren.

Die Optimierung der aus (3.2) abgeleiteten Ausdrücke erfolgt unabhängig voneinander, wobei damit begonnen wird, für eine gegebene Konfiguration der Segmentierung, die unbekannten Parameter des Szenenmodells zu bestimmen. Zur Auswertung von (3.5) werden dann die MAP-Werte der so geschätzten Verteilungsdichten unabhängig für jede Objekthypothese verwendet. Mit dem Ziel einer unüberwachten Szenensegmentierung, muss zusätzlich dazu außerdem die Anzahl der Segmentklassen $J_t$ ermittelt werden. Diese wird nicht explizit aus (3.2) abgeleitet, sondern heuristisch bestimmt. Hierauf wird in Abschnitt 4.4 näher eingegangen.

## 3.3 Beobachtungsmodelle

Die Likelihood aus (3.3) wird nun quantitativ formuliert. Für gegebene Schätzwerte $\hat{\boldsymbol{\Theta}}_t$, lässt sich der Ausdruck wie folgt umformen

$$\begin{aligned} P(\mathrm{g}_{t+1}, \mathrm{g}_t^{\mathrm{l}} | \mathbf{l}_t, \boldsymbol{\Theta}_t, \mathcal{G}_A) &= P(\mathrm{g}_{t+1} - \hat{\mathrm{g}}_{t+1}(\boldsymbol{\Theta}_t, \mathrm{g}_t) | \mathbf{l}_t, \mathrm{g}_t, \boldsymbol{\Theta}_t) \\ &\quad \times P(\mathrm{g}_t^{\mathrm{l}} - \hat{\mathrm{g}}_t^{\mathrm{l}}(\boldsymbol{\Theta}_t, \mathrm{g}_t) | \mathbf{l}_t, \mathrm{g}_t, \boldsymbol{\Theta}_t) \\ &= P(\mathbf{E}_t = \boldsymbol{\varepsilon}_t | \mathbf{l}_t, \mathrm{g}_t, \boldsymbol{\Theta}_t)\,. \end{aligned} \tag{3.7}$$

Hierbei wird angenommen, dass (i) die Wahrscheinlichkeit für das Beobachten eines Bildes durch seinen Prädiktionsfehler vollständig beschrieben ist und dass (ii), die Prädiktionsfehler in stereoskopischer und zeitlicher Richtung statistisch unabhängig sind. Unter diesen Annahmen kann die Likelihood dann mit dem Modell der Restfehlerkarte aus (3.1) formuliert werden. Der Restfehler an jeder Stelle im Bild $\mathbf{x}_{n\in\mathcal{R}}$ ist definiert als die Summe der Grauwertdifferenz zweier zeitlich und räumlich versetzter Bilder. Kann davon ausgegangen werden, dass der Restfehlerprozess bei gegebener Segmentierung stochastisch unabhängig ist, gilt

$$P(\mathbf{E}_t = \boldsymbol{\varepsilon}_t | \mathbf{l}_t, \mathrm{g}_t, \boldsymbol{\Theta}_t) = \prod_{n\in\mathcal{R}} P(\varepsilon_{n,t} | \mathbf{l}_t, \mathrm{g}_t, \boldsymbol{\Theta}_t)\,. \tag{3.8}$$

Nimmt man weiterhin an, die Likelihoodfunktion gehorche einer stationären, mittelwertfreien Normalverteilung, ergibt sich für die Komponente des bewegungsprädizierten Restfehlers

$$P(\varepsilon_{n,t}^{\mathrm{b}} | \mathbf{l}_t, \mathrm{g}_t, \boldsymbol{\Theta}_t) = \frac{1}{\sqrt{2\pi}\phi_t^{\mathrm{b}}} \exp\left[ -\frac{\Big(\mathrm{g}_{t+1}(\widehat{\boldsymbol{\chi}}_n^{\mathrm{b}}) - \mathrm{g}_t(\mathbf{x}_n)\Big)^2}{2(\phi_t^{\mathrm{b}})^2} \right] \tag{3.9}$$

und für den stereoskopischen Teil

$$P(\varepsilon_{n,t}^{\mathrm{s}} | \mathbf{l}_t, \mathrm{g}_t, \boldsymbol{\Theta}_t) = \frac{1}{\sqrt{2\pi}\phi_t^{\mathrm{s}}} \exp\left[ -\frac{\big(\mathrm{g}_t^{\mathrm{l}}(\widehat{\boldsymbol{\chi}}_n^{\mathrm{s}}) - \mathrm{g}_t(\mathbf{x}_n)\big)^2}{2(\phi_t^{\mathrm{s}})^2} \right]. \tag{3.10}$$

$\widehat{\boldsymbol{\chi}}_n^{\mathrm{b}} = \mathbf{x}_n + \mathbf{d}_t(\mathbf{x}_n, \boldsymbol{\Theta}_t)$ und $\widehat{\boldsymbol{\chi}}_n^{\mathrm{s}} = \mathbf{x}_n + \Delta_t(\mathbf{x}_n, \boldsymbol{\Theta}_t)$ drücken hier jeweils die prädizierte Position des Punktes $\mathbf{x}_n$ in (b) zeitlicher und (s) räumlicher Richtung aus. $\phi_t^{\mathrm{b}}$ und $\phi_t^{\mathrm{s}}$ beschreiben die Standardabweichung der einzelnen Restfehlerkomponenten im Bild. Die beiden Größen werden im Weiteren zusammenfassend durch $\boldsymbol{\phi}_t = (\phi_t^{\mathrm{b}}, \phi_t^{\mathrm{s}})$ beschrieben.

### 3.3.1 Behandlung von Mehrdeutigkeiten und Verdeckungen

Für Bildpunkte, die nicht eindeutig einer der initialisierten Objekthypothesen zugewiesen werden können, wird in dieser Arbeit ein zusätzliches Label eingeführt. Da dieses Label, entgegen der eigentlichen Definition eines Labelwertes in dieser Arbeit, kein real existierendes Objekt der Szene beschreibt, wird das Label als „Null-Label" bezeichnet und mit dem Wert $j = 0$ belegt. Die Definition möglicher Labelrealisationen erweitert sich somit zu $\mathcal{J} = \{0, 1, 2, \ldots, J\}$. Als Maß für die Eindeutigkeit wird dabei die Ähnlichkeit des Restfehlers der einzelnen Hypothesen verwendet. Der „Restfehler" für das Null-Label ist dabei definiert als die Differenz zwischen höchstem und geringstem Fehlerwert aller untersuchten Objekthypothesen

$$\varepsilon_{n,t}^{0} = \left|\varepsilon_{n,t}^{\{1,\ldots,J\}}\right|_{\max} - \left|\varepsilon_{n,t}^{\{1,\ldots,J\}}\right|_{\min} \qquad \forall n \in \mathcal{R}\,. \tag{3.11}$$

Für den Fall, dass keine fremdbewegten Objekte detektiert wurden, wird die Restfehlerkomponente $|\varepsilon_{n,t}|_{\min} = 0$ gesetzt. Anschaulich betrachtet, wird das Null-Label an Bildpunkt $\mathbf{x}_{n,t}$ „aktiviert", falls sich die einzelnen Objekthypothesen zu ähnlich sind, d. h. die Modellparameter für alle Hypothesen ähnliche Werte erzeugen. Der obere Teil von Abbildung 3.5 zeigt hierzu das Ergebnis einer Segmentierung in statische und fremdbewegte Bereiche relativ zur Kamera. Diese Form der Segmentierung mit der Labeldefinition $\mathcal{J} = \{0, 1\}$ wird im Weiteren auch als *Hintergrundsegmentierung* bezeichnet.

Das oben vorgestellte Modell zur Bewertung der Beobachtungen ist für alle Bildpunkte gültig, die in den jeweils versetzten Ansichten der Kameras auch existieren. Für Bereiche, die zwischen den entsprechenden Ansichten verdeckt werden, ist es erforderlich, das Modell auszusetzen, da das Restfehlerbild an solchen Stellen nicht sinnvoll definiert ist. Für die räumlich versetzten Kameras der Stereoanordnung werden Verdeckungen durch *Links-Rechts-Zuordnung* (engl. *left-right-matching*) sowie einer Überprüfung der *Ordnung* der Disparitätswerte, wie u. a. in [Brown u. a., 2003] beschrieben, detektiert. Die entsprechenden Parameter hierfür wurden in einem Trainingsschritt anhand von synthetischen und realen Bilddaten gelernt und werden innerhalb des Segmentierprozesses als konstant angenommen. Bei gegebener Segmentierung können verdeckt werdende Bereich auch eindeutig für zeitlich aufeinander folgende Bilder bestimmt werden. Eine von verschiedenen Autoren vorgeschlagene Ordnung der Labelwerte, entsprechend dem Abstand zum Kamerasystem erscheint in dieser Arbeit sinnvoll für Objektlabel, die Objekte mit begrenzter Ausdehnung beschreiben. Hier können verdeckt werdende Bereiche durch die Prädiktion des Labelfeldes, basierend auf den MAP-Schätzungen für Bewegung und Struktur, erkannt werden. In Bereichen, in denen ein bewegtes

Objekt den statischen Hintergrund verdeckt, bzw. in denen das Objekt hinter statischen Hintergrundbereichen verschwindet, wird zusätzlich die lokale Tiefeninformation ausgewertet, um mögliche Verdeckungen zu detektieren. Wird ein Punkt aufgrund der durch die Modellparameter spezifizierten Bewegung als verdeckt werdend erkannt, wird der Restfehler für die entsprechende Objekthyphothese auf einen maximal zulässigen Wert gesetzt. Mit dem so definierten Beobachtungsmodell kann, bei optimaler Schätzung von Bewegung und Struktur, eine Restfehlerkarte mit Werten nahe bei null erwartet werden.

### 3.3.2 Betrachtung stochastischer Eigenschaften

Der Restfehler ist durch einen Zufallsprozess mit Gauß'schem weißen Rauschen modelliert. Das Modell entspricht somit einem zeitinvarianten, stationären Prozess, womit die Wahrscheinlichkeiten als zeitlich und örtlich konstant angenommen werden können. Stationarität ist eine der bedeutendsten Eigenschaften stochastischer Prozesse in der Zeitreihenanalyse. Mit der Stationarität erhält man Eigenschaften, die nicht nur für einzelne Zeitpunkte bzw. Orte gelten, sondern Invarianzen über die Zeit bzw. den Ort hinweg sind. Der Zufallsprozess hat somit zu allen Zeitpunkten den gleichen Erwartungswert und die gleiche Varianz.

Unter der Annahme statistischer Unabhängigkeit des Restfehlers, kann dessen Verbundverteilung als Produkt aller univariaten Randverteilungen, wie in (3.8) beschrieben, ausgedrückt werden. Die Untersuchung der Stationaritätsannahme wurde in dieser Arbeit auf der Basis typischer Verkehrsszenarien durchgeführt. Abbildung 3.5 zeigt die empirischen univariaten Randverteilungen des Restfehlers über der Zeit am Beispiel des Hintergrundsegments, d. h. der statischen Szene relativ zum Betrachter. Die einzelnen Segmente können als relativ zur Kamera unabhängig bewegte Szenenbereiche interpretiert werden. Es zeigt sich, dass es durch die sich stetig ändernden Bildinhalte zu einer starken Schwankung der Restfehlerverteilung kommt, die nicht vernachlässigt werden kann. Die teilweise starke Fragmentierung der Bilder rührt von der mangelnden Trennschärfe der Bewegungen her. Dies stellt jedoch in keiner Weise ein Problem für das Verfahren dar, da das Null-Label nicht als eigentliches Objekt interpretierbar ist.

Durch die Auswertung der jeweiligen Bereiche hinsichtlich konsistenter Bewegung und räumlicher Kompaktheit können aus der hier gezeigten Hintergrundsegmentierung verlässlich neue Objekthypothesen generiert werden, wie im weiteren Verlauf der Arbeit gezeigt wird. Durch die Nachführung der sich zeitlich ändernden statistischen Eigenschaften der jeweiligen Segmente kann die Forderung nach Stationarität gelockert werden, woraus das Modell eines segmentweise stationären Restfehlerprozesses entsteht. Als hinreichende Näherung wird die Restfehlerver-

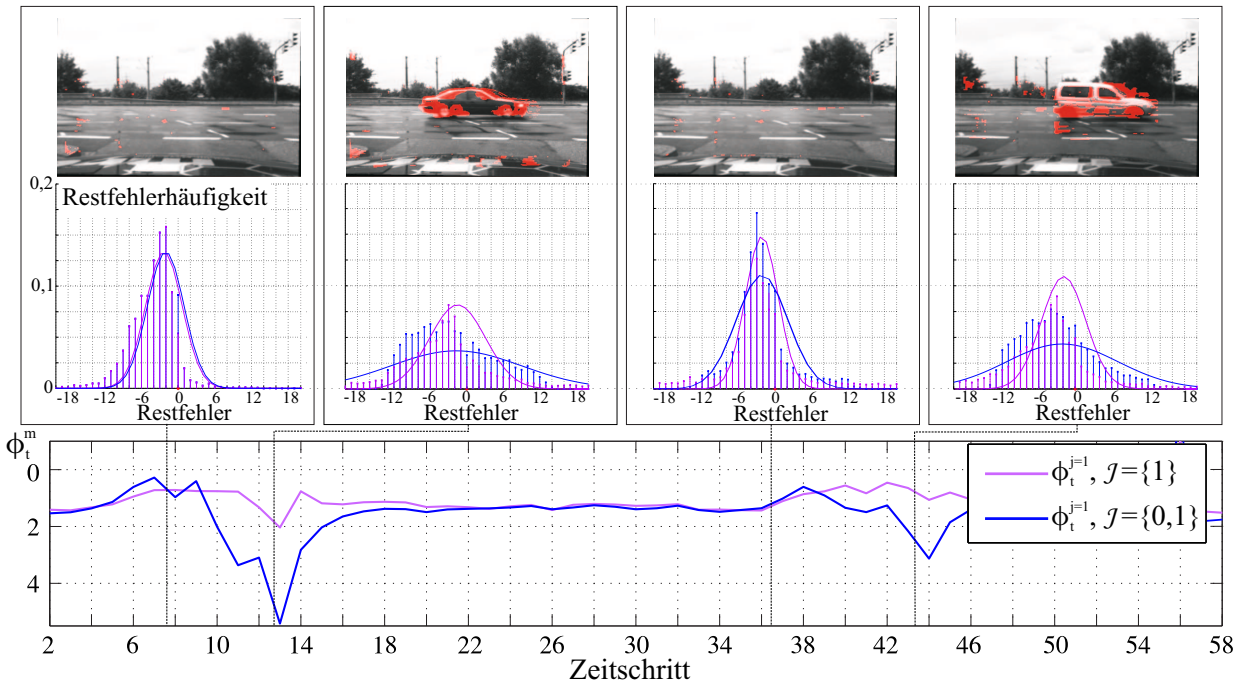



**Abbildung 3.5**: Oben: Relativ zum Betrachter unabhängig bewegte Bereiche des Bildes sind rot unterlegt (Null-Label, $j = 0$). Bereiche im Bild, die keine ausreichende Textur aufweisen, werden bei der Segmentierung ausgeblendet. Für das Hintergrundsegment sind darunter die zugehörigen empirischen univariaten Randverteilungen des Restfehlers mit entsprechender Approximation abgebildet. In blau ist die Randverteilung gezeigt unter Einbezug aller Bildpunkte. Die rote Kurve zeigt die Verteilung des Restfehlers unter Ausschluss der als fremdbewegt detektierten Bildpunkte. Unten: Entsprechende zeitliche Entwicklung der Standardabweichung des Restfehlers.

teilung durch die ML-Schätzung

$$\phi_t^j = \sqrt{\frac{1}{|\mathcal{R}^j|} \sum_{n \in \mathcal{R}^j} \varepsilon_{n,t}^2} \,, \tag{3.12}$$

für das jeweilige Segment $j \in \mathcal{J}$ approximiert. Hierbei stellt $|\mathcal{R}^j|$ die relative Häufigkeit des Labelwertes zum Zeitpunkt $t$ im gesamten Bild dar.

## 3.4 Modellierung von Vorwissen

Zur Stützung der Schätzung auf Grundlage der in (3.9) und (3.10) definierten Ausdrücke, wird durch die Berücksichtigung des zweiten Terms in (3.5) bzw. (3.6)

Vorwissen bezüglich der zu bestimmenden Größe in den Schätzprozess mit eingebracht. *Glattheit* ist in diesem Zusammenhang eine in der Bildverarbeitung vielfach verwendete Annahme und hat sich in den unterschiedlichsten Disziplinen als ausgesprochen hilfreich erwiesen [Marr, 1982; Poggio u. a., 1985; Besag, 1986; Ferrari u. a., 1995; Larsen u. a., 2006; Wedel u. a., 2007; Szeliski u. a., 2008]. Auch in dieser Arbeit wird die Glattheitsbedingung auf das Modell der Szenensegmentierung übertragen, wodurch es zu einer Bevorzugung einzelner, stückweise glatter Oberflächen kommt, die sich ähnlich im Raum bewegen. Da erwartet werden kann, dass Diskontinuitäten der Szenenbewegung und Szenenstruktur an den Objekträndern gleichermaßen auftreten, bietet sich eine gemeinsame Modellierung beider Größen in Form eines, in Abschnitt 2.4 vorgestellten, Markov-Zufallsfeldes an. Durch dieses Modell kann eine globale Glattheitsanforderung nicht nur in örtlicher, sondern zusätzlich dazu auch in zeitlicher Richtung formuliert werden.

### 3.4.1 Örtliche und zeitliche Glattheit der Segmentierung

Bei der Beschreibung der globalen Charakteristik des Segmentierprozesses $\mathbf{l}_t$ unter Verwendung eines Markov'schen Zufallsfeldes zur Modellierung stückweiser Glattheit, hilft die in Abschnitt 2.4 eingeführte Gibbs-Markov-Äquivalenz, um den zweiten Faktor in (3.5) durch den Ausdruck

$$\begin{aligned} P(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathcal{G}_A) = \mathrm{Z}_L^{-1}\exp\left[-H\left(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathrm{g}_t,\hat{\mathbf{l}}_{t-1}\right)\right], \quad \text{mit} \\ \mathrm{Z}_L = \sum_{\mathbf{l}_t\in\mathcal{F}} \exp\left[-H\left(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathrm{g}_t,\hat{\mathbf{l}}_{t-1}\right)\right] \end{aligned} \tag{3.13}$$

zu ersetzen. Die Zustandssumme $\mathrm{Z}_L$ wird durch Summation über alle Labelrealisationen im Konfigurationsraum $\mathcal{F}$ gebildet. Da man bei der Segmentierung jedoch nur an der Konfiguration mit einem ausgezeichneten Modalwert der Wahrscheinlichkeitsverteilung des Zufallsfeldes interessiert ist, entfällt diese aufwändige Berechnung hier. Das Energiefunktional im Exponenten von (3.13) setzt sich bei der Segmentierung aus den drei Termen

$$H\left(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathrm{g}_t,\hat{\mathbf{l}}_{t-1}\right) = H_L\left(\mathbf{l}_t|\mathrm{g}_t\right) + H_{L_t}\left(\mathbf{l}_t|\boldsymbol{\Theta}_t,\hat{\mathbf{l}}_{t-1}\right) + H_{L_\xi}\left(\mathbf{l}_t|\boldsymbol{\Theta}_t\right) \tag{3.14}$$

zusammen, welche jeweils eine Konsistenzbedingung bzgl. der lokalen Nachbarschaft $\mathcal{N}_n$ eines jeden Bildpunktes $\mathbf{x}_n$ formuliert:

- $H_L$ drückt die Annahme segmentweiser Glattheit des Labelfeldes aus. Da Objektgrenzen gewöhnlich auch mit hohen Gradienten im Grauwertbild $\mathrm{g}_t$ verbunden sind, wird die Glattheitsanforderung der Segmentierung hier um eine, mit Grauwertkanten übereinstimmende, Segmentberandung erweitert.

- ◇ Neben der örtlichen, wird in dieser Arbeit außerdem auch die zeitliche Glattheit bezüglich der vorangegangenen Szenensegmentierung berücksichtigt. Dies geschieht durch den zweiten Term $H_{L_t}$, der die Erwartung einer zeitlich konsistenten Segmentierung entlang der Bewegungstrajektorie des Objekts formuliert.

- ◇ Aufgrund der Komplexität natürlicher Verkehrsumgebungen und der darin vorkommenden Bewegungen hat sich die Segmentierung der Szene, basierend auf dem Bewegungsmerkmal allein, in einigen Situationen als unzureichend erwiesen. In den meisten dieser Fälle, führt die mangelnde Trennschärfe der einzelnen Bewegungsprofile zu einer starken Übersegmentierung der Szene. Aus diesem Grund wird bei den Hypothesen fremdbewegter Objekte das Bewegungsmerkmal selbst noch um die räumliche Ausdehnung des jeweiligen Objekts im Raum erweitert, was durch $H_{L_\xi}$ in obiger Gleichung ausgedrückt wird.

### Örtliche Glattheit der Segmentierung

Die Annahme stückweiser Glattheit des Labelfeldes in örtlicher Richtung wird durch das Energiefunktional

$$H_L\left(\mathbf{l}_t|\mathrm{g}_t\right) = \sum_{c\in\mathcal{C}} \lambda_1 V_c\left(\mathbf{l}_t, \mathrm{g}_t\right) \tag{3.15}$$

beschrieben. In dieser Arbeit wird das Cliquenpotential aller benachbarten Bildpunkte als zweielementige Clique $c = \langle n, u\rangle$ modelliert, wobei $|\mathbf{x}_n - \mathbf{x}_u| = 1$ ist. Diese Definition einer Clique ist gleichbedeutend mit einem Nachbarschaftssystem erster Ordnung, das aus den direkten Nachbarn der Labelvariable $l_t(\mathbf{x}_n) = l_{n,t} \in \mathbf{l}_t$ in horizontaler und vertikaler Richtung besteht. Das Potential der so definierten Clique

$$V_{\langle n,u\rangle}\left(l_{n,t}, l_{u,t}, \mathrm{g}_t\right) = \begin{cases} 0 & \text{falls } \; l_{n,t} = l_{u,t} \\ \min\left(s_{\langle n,u\rangle}(\mathrm{g}_t), T_L\right) & \text{sonst}\,, \end{cases} \tag{3.16}$$

kann als Erweiterung des in (2.30) vorgestellten Potts-Modells verstanden werden. Durch dieses Modell werden gleiche Nachbarschaftslabel bevorzugt und somit die Glattheitsbedingung in einer bestimmten Nachbarschaft favorisiert. Bei gleichen Labelwerten $l_n$ und $l_{u\in\mathcal{N}_n}$, erzeugt der Glattheitsterm ein Energiepotential von 0. Bei unterschiedlichen Labelwerten ist die Energie definiert durch die Funktion $s_{\langle n,u\rangle}(\mathrm{g}_t)$, welche im einfachsten Fall durch eine geeignete Konstante vorgegeben werden kann. In dieser Arbeit soll das Glattheitskriterium um die Eigenschaft

der Grauwertkonsistenz [Poggio u. a., 1985; Marr, 1982] innerhalb eines bewegten Objekts erweitert werden, was durch

$$s_{\langle n,u\rangle}(\mathrm{g}_t) = \left[1 + \frac{\lambda_\mathrm{g}\,|\mathrm{g}_{n,t} - \mathrm{g}_{u,t}|}{\sigma_\mathrm{g}}\right]^{-1} \tag{3.17}$$

erreicht wird. $T_L \in \mathbb{R}_+$ begrenzt das Potential auf einen maximal zulässigen Wert. $\lambda_1, \lambda_\mathrm{g} \in \mathbb{R}_+$ stellen Regularisierungsparameter dar, die den Einfluss der Glattheitsanforderung auf die Segmentierung bestimmen. $\sigma_\mathrm{g}$ definiert die Steilheit der Gewichtsfunktion und wird, wie die anderen Parameter auch, in einem Trainingschritt ermittelt. Durch (3.17) wird unterschiedliches Labeling benachbarter Punkte stärker bestraft, wenn die entsprechenden Bildpunkte ähnliche Grauwerte haben. Dadurch wird berücksichtigt, dass Objekte oftmals als zusammenhängende Bereiche ähnlicher Farbe im Bild gegeben sind.

### Zeitliche Glattheit der Segmentierung

Der zweite Term in (3.14) formuliert die Erwartung einer zeitlich konsistenten Segmentierung entlang der Bewegungstrajektorie des jeweiligen Objekts. Die kausale Abhängigkeit des Zufallsfeldes lässt sich durch einfache Potentiale einelementiger Cliquen $c = \langle n\rangle$ einbringen. Durch diese Einercliquen wird ein Bezug zur Schätzung des vorherigen, als konstant anzunehmenden Labelfeldes $\hat{\mathbf{l}}_{t-1}$ aufgebaut. Das entsprechende Energiefunktional ist dann folgendermaßen definiert

$$\begin{aligned} H_{L_t}\left(\mathbf{l}_t \,|\, \boldsymbol{\Theta}_t, \hat{\mathbf{l}}_{t-1}\right) &= \sum_{\langle n\rangle \in \mathcal{C}} \lambda_2\, V_{\langle n\rangle}\left(\mathbf{l}_t, \boldsymbol{\Theta}_t, \hat{\mathbf{l}}_{t-1}\right), \quad \text{mit} \\ V_{\langle n\rangle}\left(\mathbf{l}_t, \boldsymbol{\Theta}_t, \hat{\mathbf{l}}_{t-1}\right) &= \begin{cases} 0 & \text{falls } \; l_{n,t} = \hat{l}_{\widehat{\boldsymbol{\chi}}_t, t-1} \\ 1 & \text{sonst}\,. \end{cases} \end{aligned} \tag{3.18}$$

$\widehat{\boldsymbol{\chi}}_t$ bezeichnet hier die rückprädizierte Position des Punktes $\mathbf{x}_n$ im vorherigen Bild, d. h. $\widehat{\boldsymbol{\chi}}_t = \widehat{\boldsymbol{\chi}}_{t-1}(\mathbf{x}_n, \boldsymbol{\Theta}_t) = \mathbf{x}_n - \mathbf{d}_t(\mathbf{x}_n, \boldsymbol{\Theta}_t)$. $\lambda_2 \in \mathbb{R}_+$ ist der entsprechende Regularisierungsparameter, der den Einfluss der zeitlichen Konsistenzforderung auf die Segmentierung steuert. Sinnvollerweise wird die zeitliche Konsistenzbedingung in verdeckt werdenden Bereichen ausgesetzt.

### Berücksichtigung der Objektausdehnung

Die Annahme einer räumlich begrenzten Ausdehnung von relativ zum Hintergrundsegment unabhängig bewegten Objekten wird ebenfalls mit Hilfe von Einer-

cliquen modelliert und bei der Segmentierung durch das Funktional

$$
\begin{aligned}
H_{L_{\boldsymbol{\xi}}}\left(\mathbf{l}_t|\boldsymbol{\Theta}_t\right) &= \sum_{\langle n\rangle\in\mathcal{C}} \lambda_{\boldsymbol{\xi}}\, V_{\langle n\rangle}\left(\mathbf{l}_t,\mathbf{z}_t,\boldsymbol{\xi}_t\right), \quad \text{mit} \\
V_{\langle n\rangle}\left(\mathbf{l}_t,\mathbf{z}_t,\boldsymbol{\xi}_t\right) &= \min\left((\mathbf{X}_{n,t}-\mathbf{M}_t^{\mathrm{T}})^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\Sigma}_t)(\mathbf{X}_{n,t}-\mathbf{M}_t^{\mathrm{T}}), T_{\boldsymbol{\xi}}\right)
\end{aligned}
\tag{3.19}
$$

berücksichtigt. $\mathrm{diag}(\boldsymbol{\Sigma}_t)$ drückt hier eine Diagonalmatrix mit den Ausdehnungskomponenten auf der Hauptdiagonalen aus. $\mathbf{X}_{n,t}$ beschreibt einen beliebigen Szenenpunkt in der Welt. $\lambda_{\boldsymbol{\xi}} \in \mathbb{R}_+$ steuert, entsprechend den anderen Regularisierungsparametern, den Einfluss der Objektausdehnung auf das Segmentierergebnis. Besondere Sorgfalt muss hier darauf gelegt werden, dass der Eintrag von (3.19) zur Gesamtenergie nicht zu groß wird, da ansonsten der bewegungssensitive Anteil der Segmentierung immer weiter in den Hintergrund rückt. Gerade die Allgemeingültigkeit des Bewegungsmerkmals soll bei der Segmentierung aber erhalten bleiben. Deshalb wird der Wert von $H_{L_{\boldsymbol{\xi}}}$ auf einen maximal zulässigen Wert $T_{\boldsymbol{\xi}} \in \mathbb{R}_+$ begrenzt. Im unteren Teil von Abbildung 3.6 ist beispielhaft der Einfluss von (3.19) auf das Segmentierergebnis für anwachsende Werte von $\lambda_{\boldsymbol{\xi}}$ illustriert.

### Training der MRF-Parameter des Segmentierungsmodells

Um eine Segmentierung auf der Basis des oben vorgestellten Modells durchführen zu können, müssen die MRF-Parameter $\boldsymbol{\varphi}_L = (\lambda_1, \lambda_2, \lambda_{\mathrm{g}}, \lambda_{\boldsymbol{\xi}}, \sigma_{\mathrm{g}}, T_L, T_{\boldsymbol{\xi}})$ der Segmentierungsmodells bekannt sein oder aus den Daten selbst geschätzt werden. Die Wahl der Werte hat dabei einen entscheidenden Einfluss auf das Segmentierergebnis. In Abbildung 3.6 ist hierzu der Einfluss einzelner Modellparameter auf das Segmentierergebnis einer Verkehrsszene mit einem fremdbewegten Objekt ($|\mathcal{J}| = 3$) gezeigt.

Im Rahmen dieser Arbeit wurden die Parameter in einem Trainingsschritt auf der Basis realer und synthetisch generierter Bilder ermittelt. Als Trainingsverfahren wurde ein Ansatz, ähnlich dem in Anhang A.1 beschriebenen Verfahrens des simulierten Abkühlens (engl. *simulated annealing*), verwendet. Zur Bewertung der Segmentierung auf Basis einer zufällig gewählten Parameterkonfiguration $\boldsymbol{\varphi}_L^k, k = (1,\ldots,K)$, wurde die Summe der als falsch klassifizierten Bildpunkte $C_L$ bezüglich der bekannten, wahren Segmentierung gewählt. Abbildung 3.7 zeigt beispielhaft den Iterationsprozess für zwei Trainingsbilder. Das Training der Parameter wurde auf insgesamt 50 Bildern durchgeführt, wobei in den meisten Fällen eine Anzahl von $K = 4000$ Iterationsschritten ausreichte, um einen Segmentierungsfehler $C_L < 5\%$ zu gewährleisten.

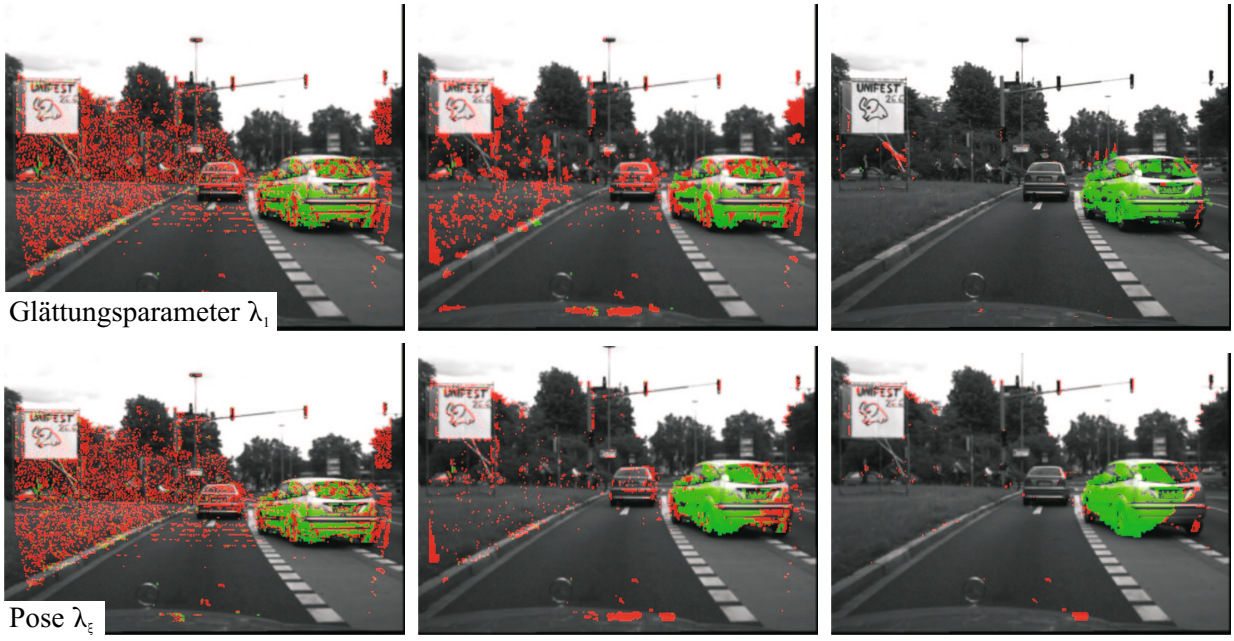



**Abbildung 3.6**: Gezeigt ist die Segmentierung für unterschiedliche Werte der Glättungsvariablen $\lambda_1$ und $\lambda_\xi$. Der Wert des jeweiligen Parameters beginnt dabei mit einem Betrag von null und steigt von links nach rechts stetig an. Grün unterlegte Bildpunkte ($\mathcal{X}^{j=2}$) sind der Objekthypothese des unabhängig bewegten Objekts zugeordnet. Rote Bereiche ($\mathcal{X}^{j=0}$) können nicht eindeutig einer der beiden Hypothesen zugewiesen werden und tragen somit das Null-Label. Die statischen Punkte relativ zur Kamera sind farblich nicht gekennzeichnet.

### 3.4.2 Örtliche und zeitliche Glattheit der Szenenstruktur

Die oben getroffenen Annahmen bzgl. der Szenensegmentierung sind sehr eng mit den zu erwartenden Eigenschaften der Szenenstruktur verbunden. So kann angenommen werden, dass ein gleichförmige bewegtes Objekt in der Szene auch über eine kontinuierliche, stückweise glatte Oberfläche verfügt. An Objektgrenzen muss diese Bedingung in den meisten Fällen ausgesetzt werden, was für bewegte Objekte durch die Segmentierung erreicht werden kann. Analog zur Segmentierung werden die Glattheitsannahmen bzgl. der dichten Szenenstruktur $\mathbf{Z}_t \in \mathbf{\Theta}_t$ in Form eines Markov-Zufallsfeldes modelliert. Der zweite Term in (3.6) wird dabei durch folgende Gibbs-Verteilung beschrieben

$$P(\mathbf{z}_t|\mathbf{l}_t,\mathcal{G}_A) = \mathrm{Z}_Z^{-1} \exp\left[-H\left(\mathbf{z}_t|\mathbf{l}_t,\hat{\mathbf{l}}_{t-1},\hat{\mathbf{\Theta}}_{t-1}\right)\right] \quad , \quad \text{mit}$$
$$\mathrm{Z}_Z = \sum_{\mathbf{z}_t \in \mathcal{F}} \exp\left[-H\left(\mathbf{z}_t|\mathbf{l}_t,\hat{\mathbf{l}}_{t-1},\hat{\mathbf{\Theta}}_{t-1}\right)\right] . \tag{3.20}$$

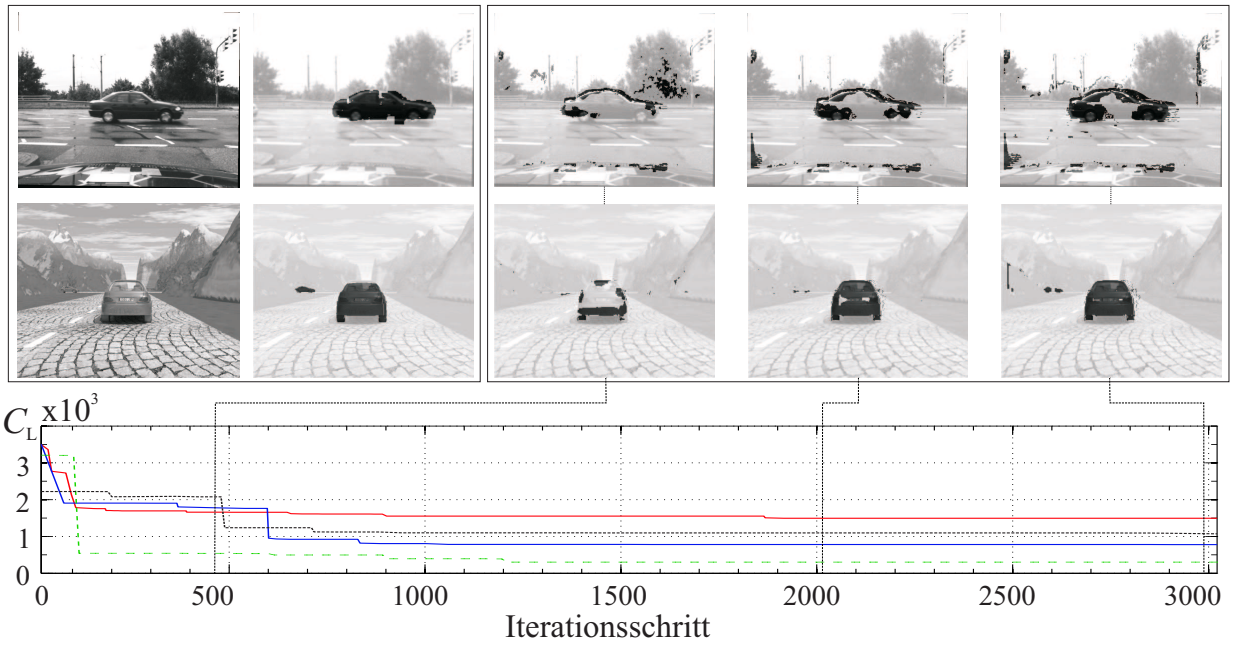



**Abbildung 3.7**: Oben: Die erste Spalte zeigt zwei Trainingsbilder aus der Datenbank. In der zweiten Spalte ist die dazu gehörende, optimale Segmentierung dargestellt. Bildpunkte, die dem statischen Hintergrund zugeordnet sind, werden bei dieser Darstellung überblendet. Bewegte Bildbereiche sind hervorgehoben. Die Folgebilder zeigen jeweils Momentaufnahmen des Trainingsprozesses für eine zufällig gewählte Parameterkonfiguration $\boldsymbol{\varphi}_L^k$. Unten: Die zeitliche Entwicklung des Fehlers $C_L$. Eine spezifische Konfiguration wird beibehalten, falls die entsprechende Segmentierung zu einer Minimierung des Fehlers $C_L$ führt.

Durch eine geeignete Modellierung des Energiefunktionals

$$H\left(\mathbf{z}_t|\mathbf{l}_t,\hat{\mathbf{l}}_{t-1},\hat{\boldsymbol{\Theta}}_{t-1}\right) = H_Z\left(\mathbf{z}_t|\mathbf{l}_t\right) + H_{Z_t}\left(\mathbf{z}_t|\hat{\mathbf{l}}_{t-1},\hat{\boldsymbol{\Theta}}_{t-1}\right) \tag{3.21}$$

kann die Glattheitsbedingung auch auf den Rekonstruktionsprozess erweitert werden. Hierbei drückt $H_Z$ die die Erwartung einer stückweise glatten Szenenoberfläche aus, während der zweite Term Vorwissen aus zeitlich zurückliegenden Schätzungen bei der Bestimmung der aktuellen Szenenstruktur mit einbezieht.

### Örtliche Glattheit der Szenenstruktur

Der erste Term in (3.21) bewertet den Unterschied zwischen den Tiefenwerten $z_{n,t}$ und $z_{u\in\mathcal{N}_n,t}$ benachbarter Bildpunkte. Hierfür werden ebenfalls Paarcliquen

$c = \langle n,u \rangle$ verwendet, d. h.

$$H_Z\left(\mathbf{z}_t|\mathbf{l}_t\right) = \sum_{\langle n,u\rangle \in \mathcal{C}} \lambda_3 \, V_{\langle n,u\rangle}\left(\mathbf{z}_t, \mathbf{l}_t\right) . \tag{3.22}$$

Neben der Szenentiefe $\mathbf{z}_t$ selbst, wird außerdem die aktuelle Segmentierung dazu genutzt, um die Glattheitsannahme an den aktuell geschätzten Segmentgrenzen auszusetzen. Die entsprechende Potentialfunktion ist dann definiert durch

$$V_{\langle n,u\rangle}\left(\mathbf{z}_t, \mathbf{l}_t\right) = \begin{cases} \min\left(\|z_{n,t} - z_{u,t}\|^w, T_Z\right) & \text{falls } l_{n,t} = l_{u,t} \\ T_z & \text{sonst .} \end{cases} \tag{3.23}$$

Die Werte der Potentialfunktion entsprechen einer Potenz $w$ der euklidischen Norm $\|\cdot\|$, welche mit zunehmendem Tiefenunterschied der Punkte monoton ansteigen. Bei diesem Modell müssen jedoch die an Objektkanten zu erwartenden Tiefensprünge berücksichtigt werden, die ansonsten zu unzulässig hohen Fehlereinträgen durch das Modell selbst führen würden [Poggio u. a., 1985]. Energiefunktionale, die diesem Umstand Rechnung tragen, werden als diskontinuitäts-erhaltend (engl. *discontinuity preserving*) bezeichnet und meist durch die Erweiterung der Potentialfunktion um einen Schwellwertes $T_Z$ umgesetzt [Black u. Rangarajan, 1996]. Übersteigt die Tiefendifferenz diesen Schwellwert, steigt die Fehlerenergie nicht weiter an, was eine unerwünschte Glättung bestehender Tiefensprünge begrenzt. Die Modellierung von Diskontinuitäten kann hier auch explizit durch einen Linienprozess (engl. *line process*) [Geman u. Geman, 1984] erfolgen, oder auch – unter Einbezug der Bilddaten selbst – mit dem Auftreten von Kanten im Intensitätsbild bzw. der Grauwerthomogenität innerhalb von Bereichen ähnlicher Tiefe [Fua, 1993; Boykov u. a., 2001] gleich gesetzt werden.

### Zeitliche Glattheit der Szenenstruktur

Zusätzlich zu den örtlichen Bindungen werden in dieser Arbeit weiterhin die starken statistischen Bindungen der Szenenstruktur in zeitlicher Richtung berücksichtigt. Diese Bindungen werden durch einelementige Cliquen modelliert, womit sich für den zweiten Term in (3.21) der folgende Ausdruck ergibt:

$$H_{Z_t}\left(\mathbf{z}_t|\hat{\mathbf{l}}_{t-1}, \hat{\boldsymbol{\Theta}}_{t-1}\right) = \sum_{\langle n\rangle \in \mathcal{C}} \lambda_4 \, V_{\langle n\rangle}\left(\mathbf{z}_t, \hat{\mathbf{l}}_{t-1}, \hat{\boldsymbol{\Theta}}_{t-1}\right) . \tag{3.24}$$

Die enthaltene Potentialfunktion

$$V_{\langle n\rangle}\left(\mathbf{z}_t, \hat{\mathbf{l}}_{t-1}, \hat{\boldsymbol{\Theta}}_{t-1}\right) = \min\left(\|z_{n,t} - \hat{z}_{n,t|t-1}\|^w, T_Z\right) \tag{3.25}$$

bewertet hierbei die Ähnlichkeit der Werte von aktuell geschätzter und erwarteter Strukturkomponente aus dem letzten Zeitschritt. Unter der Annahme starrer Objektbewegung und basierend auf den Schätzwerten $\hat{\mathbf{l}}_{t-1}$, $\hat{\boldsymbol{\theta}}_{t-1}$ und $\hat{\mathbf{z}}_{t-1}$ kann mit Hilfe der inversen Projektionsgleichung $\Pi^{-1}$ die erwartete 3D Position jedes Punktes

$$\hat{\mathbf{X}}_{m,t} = \mathbf{R}\left(\boldsymbol{\omega}_t^{\hat{l}_{m,t-1}}\right)\Pi^{-1}(\mathbf{x}_{m,t-1}, \hat{z}_{m,t-1}) + \mathbf{t}_t^{\hat{l}_{m,t-1}} \tag{3.26}$$

aus dem zeitlich zurückliegenden Punkt bestimmt werden. Die einzelnen Komponenten $\hat{z}_{n,t|t-1}$ werden dann aus der resultierenden bewegungsprädizierten Tiefenkarte extrahiert. Der Wertebereich der Potentialfunktion wird ebenfalls durch den Schwellwert $T_Z$ nach oben beschränkt. An Punkten, für die kein solcher Prädiktionswert vorliegt, wird die Potentialfunktion ausgesetzt. Dieser Fall tritt vor allem zu Beginn der Szenensegmentierung auf, da in diesem Stadium eine noch ungenaue Bewegungsschätzung zu einer Segmentierung führt, die in weiten Bereichen des Bildes keine eindeutige Zuordnung der Bildpunkte erlaubt. Als Konsequenz daraus wird vermehrt das Null-Label aktiviert. Da für solche Punkte kein eindeutiges Bewegungsprofil identifiziert werden kann, entfällt hier auch die Prädiktion der Szenentiefe. Das Null-Label wird außerdem dafür genutzt, um den Prädiktionsterm in Bereichen im Bild auszusetzen, in denen die Rekonstruktion zu schlechten Ergebnissen führt. Durch die Verschlechterung steigt der Restfehler in diesen Bereichen für alle getesteten Hypothesen gleichermaßen an, was zur Folge hat, dass dort das Null-Label zugewiesen wird. Abbildung 3.8 zeigt hierzu die zeitliche Entwicklung einer Tiefenkarte, die nur die bewegungsprädizierten Strukturkomponenten aus dem jeweils vorhergehenden Zeitschritt berücksichtigt. In der Praxis hat sich gezeigt, dass die Verwendung von (3.24) zu keiner wesentlichen Verbesserung der Schätzergebnisse bei der Szenenrekonstruktion selbst führt. Jedoch können durch deren Einführung unplausible Tiefenwerte bereits im Vorfeld eliminiert werden, wodurch der Rechenaufwand des Verfahrens um bis zu 20% bzgl. der ursprünglichen Rechenzeit reduziert werden konnte.

### Training der MRF-Parameter des Strukturmodells

Die MRF-Parameter des Strukturmodells $\boldsymbol{\varphi}_Z = (\lambda_3, \lambda_4, T_Z)$ wurden auf die gleiche Weise wie die MRF-Parameter der Segmentierung in einem Parametertraining ermittelt. Neben den Modellparametern selbst wurden zusätzlich noch eine Reihe von implementierungsspezifischen Parametern optimiert, auf die in dieser Arbeit nicht näher eingegangen werden soll. Aufgrund der vorhandenen Referenzdaten in

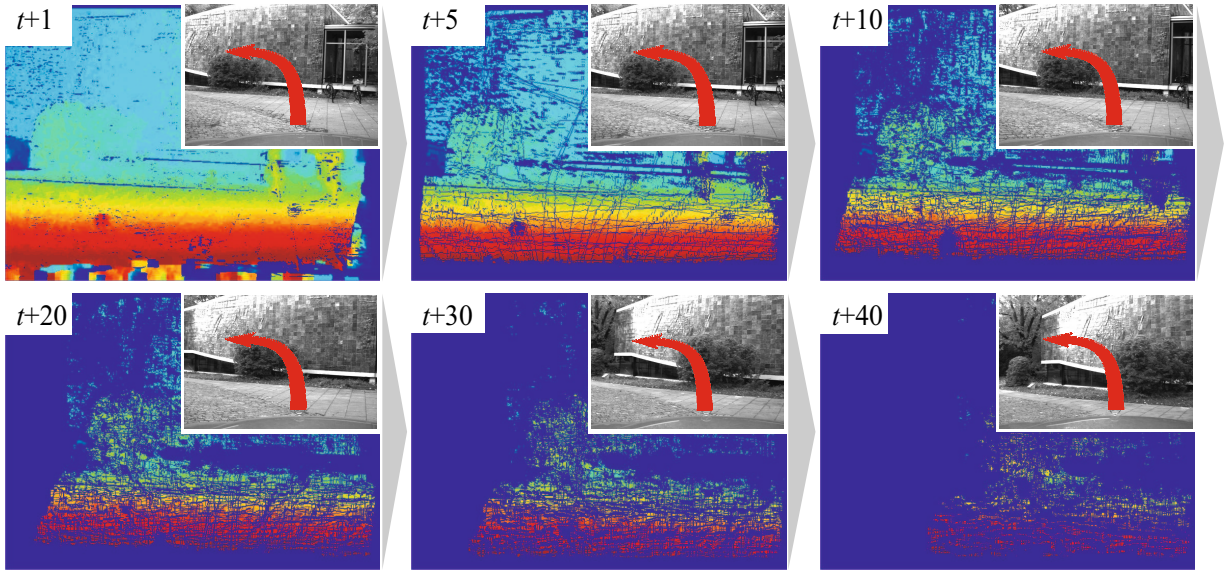



**Abbildung 3.8**: Die zeitliche Entwicklung der bewegungsprädizierten Strukturkomponenten. Zum Zeitpunkt $t = 0$ wird das Rekonstruktionsergebnis „eingefroren" und über die Zeit weiterprädiziert. Aufgrund der Eigenbewegung des Fahrzeugs (symbolisiert durch den roter Pfeil im Bild oben rechts), wandert die Tiefenkarte nach unten rechts aus dem Bild heraus. Bildpunkte, die bei der Segmentierung dem Null-Label zugewiesen werden, verschwinden ebenfalls aus der Karte.

Form von dichten Disparitätsbildern wurde als Gütemaß der arithmetische Mittelwert

$$C_\Delta = \frac{1}{N} \sum_{n \in \mathcal{R}} |\Delta_{n,\mathrm{GT}} - \Delta_{n,\mathrm{EST}}| \tag{3.27}$$

des Rekonstruktionsfehlers gewählt. Index „GT" steht hier für „Ground Truth"[8], „EST" für „Estimation". Abbildung 3.9 zeigt beispielhaft einige Momentaufnahmen aus dem Parametertraining für verschieden Trainingsbilder. Das Training wurde auf der Basis von 70 Stereobildpaaren, die unterschiedlichste Szenen abbilden, durchgeführt.

[8] „Ground Truth" kommt aus dem Bereich der Fernerkundung und Kartografie und bezeichnet direkt durch Geländeerkundung *am Boden* aufgenommene Informationen, die zur Analyse und Bewertung von Luftaufnahmen, Satellitenbildern, etc. genutzt werden. Die Bezeichnung hat sich auch in anderen Bereichen der Bildverarbeitung etabliert und wird allgemein zur Bezeichnung bekannter Referenzdaten verwendet, die zur Validierung von Bildverarbeitungsalgorithmen dienen.

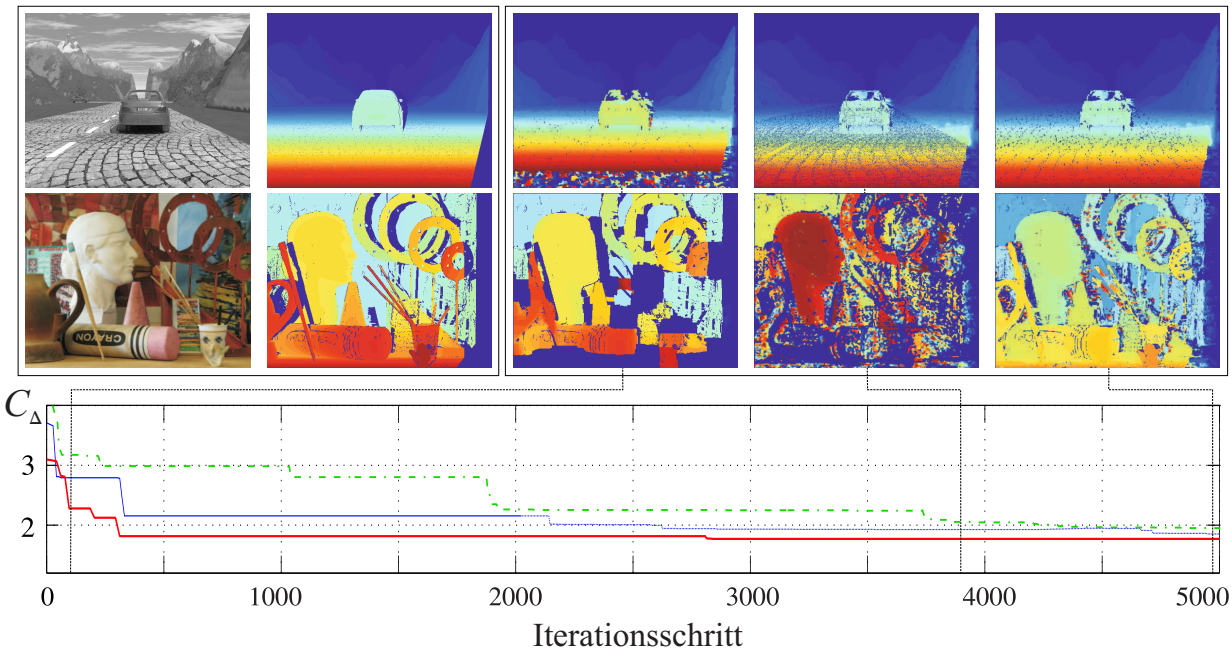



**Abbildung 3.9**: Oben: Die erste Spalte zeigt eine Auswahl an Trainingsbildern, die für das Training der MRF-Parameter des Strukturmodells verwendet wurden. In der zweiten Spalte sind die bekannten Tiefenwerte der einzelnen Szenenpunkte in Form einer Tiefenkarte dargestellt. Die Farbe codiert die entsprechende Szenentiefe: rote Farben symbolisieren Nähe; mit anwachsender Entfernung gehen die Farben zunehmend ins Blaue über. Die jeweiligen Bildfolgen in horizontaler Richtung zeigen Momentaufnahmen des Trainings. Unten: Der Fehler im Bild für eine spezielle Parameterkonfiguration $\boldsymbol{\varphi}_Z^k$ über der Zeit.

## 3.5 Gesamtmodell der dichten Szenensegmentierung

Das oben vorgestellte Modell erlaubt eine dichte Szenensegmentierung und stellt dabei Glattheitsanforderungen in örtlicher und zeitlicher Richtung an die Rekonstruktion und die Segmentierung. Die Optimierung der einzelnen Größen wird in dieser Arbeit ortsdiskret durchgeführt, d. h. die Auflösung der resultierenden Segmentierung ist durch die Auflösung des Bildgitters beschränkt. Auf die Möglichkeit einer Interpolation der sich hieraus ergebenden Werte, wie z. B. in [Scharstein u. Szeliski, 2001] für die Bestimmung subpixelgenauer Disparitätswerte erläutert, wird in dieser Arbeit verzichtet. Bei dem hier verwendeten globalen Optimierungsverfahren wird eine optimale Zuordnung der Bildpunkte im Referenz- und Suchbild durch Maximierung des Ausdrucks in (3.6) erreicht. Hierfür wurde ein Modell vorgestellt, welches bestimmte Glattheitsanforderungen in örtlicher und zeitlicher Richtung an den Rekonstruktionsprozess stellt. Mit Hilfe

von (3.10) und (3.20) kann (3.6) dann, entsprechend den Ausführungen in Abschnitt 2.5.1, folgendermaßen umformuliert werden:

$$\begin{aligned}\hat{\mathbf{z}}_{t,\mathrm{MAP}} &= \arg\max_{\mathbf{z}_t} \left\{ \prod_{n\in\mathcal{R}} P(\boldsymbol{\varepsilon}^{\mathrm{s}}_{n,t}|\mathbf{z}_t,\mathbf{l}_t,\mathrm{g}_t)\ \exp\left[-H\left(\mathbf{z}_t|\mathbf{l}_t,\hat{\mathbf{l}}_{t-1}\hat{\boldsymbol{\Theta}}_{t-1}\right)\right]\right\} \\ &= \arg\max_{\mathbf{z}_t} \left\{ \exp\left[\sum_{n\in\mathcal{R}} -\mathrm{D}^{\mathrm{s}}(z_{n,t},\mathrm{g}_t) - H\left(\mathbf{z}_t|\mathbf{l}_t,\hat{\mathbf{l}}_{t-1},\hat{\boldsymbol{\Theta}}_{t-1}\right)\right]\right\}.\end{aligned} \tag{3.28}$$

Das Auffinden des MAP-Schätzers kann somit dem mathematisch einfacher zu bestimmenden Minimum des Energiefunktionals

$$\mathrm{E}(\mathbf{z}_t) = \sum_{n\in\mathcal{R}} -\mathrm{D}^{\mathrm{s}}(z_{n,t},\mathrm{g}_t) - H\left(\mathbf{z}_t|\mathbf{l}_t,\hat{\mathbf{l}}_{t-1},\hat{\boldsymbol{\Theta}}_{t-1}\right) \tag{3.29}$$

gleichgesetzt werden. Bei der Rekonstruktion wird jedem Bildpunkt ein Wert $z_{n,t} \in \mathcal{Z}$ aus der Menge $\mathcal{Z} = \{0,\ldots,\mathrm{T}\} \in \mathbb{N}$ möglicher Tiefenwerte zugewiesen, wobei der erste Term in (3.29) die Ähnlichkeit der Daten in einer lokalen Umgebung im Bild bewertet. Der zweite Term tritt in Bereichen in Erscheinung, die schwach bzw. gleichförmig texturiert sind, da hier der Datenterm nur über eine begrenzte Beschreibungskraft verfügt [Aschwanden u. Guggenbuhl, 1993; Zabih u. J. Woodfill, 1994; Bhat u. Nayar, 1998]. Neben einer Parametrierung der Stereorekonstruktion im 3D Konfigurationsraum $\mathcal{Z}$ [Larsen u. a., 2006], wurde alternativ dazu auch eine Rekonstruktion aus dem 2D Disparitätsraum $\mathcal{P}$ untersucht. Die entsprechenden Annahmen und Beschreibungen zur Modellierung der statistischen Bindungen wurden hier aus dem oben vorgestellten Glattheitsmodell übernommen und auf die Disparitätsschätzung übertragen. Beide Parametrierungen führten zu ähnlichen Ergebnissen, wobei die direkte Stereorekonstruktion aufgrund eines größeren Konfigurationsraums rechnerisch aufwändiger ist.

Liegen die Schätzungen der aktuellen Szenenstruktur und Objektparameter vor, kann die eigentliche Szenensegmentierung durchgeführt werden. Um den Ausdruck in (3.5) effizient mit Graphenschnittverfahren lösen zu können, wird das Optimierungsproblem – ähnlich wie bei der Schätzung der Strukturvariable – mit Hilfe von (3.8) und (3.13) folgendermaßen umformuliert

$$\begin{aligned}\hat{\mathbf{l}}_{t,\mathrm{MAP}} &= \arg\max_{\mathbf{l}_t} \left\{ \prod_{n\in\mathcal{R}} P(\boldsymbol{\varepsilon}_{n,t}|\mathbf{l}_t,\mathrm{g}_t,\boldsymbol{\Theta}_t)\ \exp\left[-H\left(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathrm{g}_t,\hat{\mathbf{l}}_{t-1}\right)\right]\right\} \\ &= \arg\max_{\mathbf{l}_t} \left\{ \exp\left[\sum_{n\in\mathcal{R}} -\mathrm{D}(l_{n,t},\mathrm{g}_t,\boldsymbol{\Theta}_t) - H\left(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathrm{g}_t,\hat{\mathbf{l}}_{t-1}\right)\right]\right\}.\end{aligned} \tag{3.30}$$

Eine Minimierung des Terms im Exponenten dieser Gleichung erfolgt mit Hilfe globaler Optimierungsalgorithmen. Hierbei können Verfahren, wie z. B. simuliertes Abkühlen (engl. *simulated annealing*) [Geman u. Geman, 1984] oder dynamisches Programmieren (engl. *dynamic programming*) [Birchfield u. Tomasi, 1998], eingesetzt werden. Aktuellere Vertreter hierzu sind u. a. Graphenschnittverfahren [Roy u. Cox, 1998; Veksler, 1999; Kolmogorov u. Zabih, 2001] oder Belief-Propagation-Ansätze [Felzenszwalb u. Huttenlocher, 2006; Larsen u. a., 2006]. Für die Bestimmung der Segmentierung und Szenenstruktur wurden in dieser Arbeit die beiden letzt genannten Ansätze implementiert und miteinander verglichen. Bei der Implementierung des Datenterms D wurde zusätzlich noch ein Parameter $T_{\mathrm{D}}$ eingeführt, um die Kosten der Beobachtungen auf einen maximal zulässigen Betrag zu beschränken. Die resultierenden Energieterme

$$\begin{aligned} \mathrm{D}^{\mathrm{s}}(z_{n,t}, \mathrm{g}_t) &= \min\left((\phi_t^{\mathrm{s}})^{-1}\varepsilon_{n,t}^{\mathrm{s}}, T_{\mathrm{D}}\right), \text{ bzw.} \\ \mathrm{D}^{\mathrm{b}}(\mathrm{l}_{n,t}, \mathrm{g}_t, \boldsymbol{\Theta}_t) &= \min\left((\phi_t^{\mathrm{b}})^{-1}\varepsilon_{n,t}^{\mathrm{b}}, T_{\mathrm{D}}\right) \end{aligned} \tag{3.31}$$

machen die Segmentierung und die Rekonstruktion robust gegenüber Verletzungen der Annahme eines konstanten Grauwerts korrespondierender Bildpunkte, die durch Verdeckungen einzelner Bildbereiche oder auch spekulare Effekte auftreten. Es hat sich gezeigt, dass beide Verfahren bei ähnlichem Rechenaufwand zu beinahe identischen Ergebnissen – sowohl bei der Rekonstruktion als auch bei der Segmentierung – führen. Die entsprechenden Parameter der einzelnen Verfahren wurden ebenfalls in einem Trainingsschritt anhand einer repräsentativen Datenbank ermittelt.

## 3.6 Zusammenfassung

Das Kapitel befasst sich mit der formalen Beschreibung des verwendeten Objekt- und Szenenmodells. Zur Beschreibung des direkten Fahrzeugumfeldes wird das geometrische Modell einer Ebene eingeführt. Die sich hieraus ergebende Fahrbahnschätzung wird genutzt, um eine Vorsegmentierung der Szene durchzuführen. Für die verbleibenden Bereiche wird die Szenensegmentierung als Schätzproblem im Bayes'schen Sinne formuliert. Die so entstandene Schätzaufgabe besteht, neben der Bestimmung der Segmentierung und Szenenstruktur, außerdem noch aus der Bestimmung der Objekteigenschaften 3D Bewegung sowie der Lage und Ausdehnung des jeweiligen Objekts relativ zur Kamera. Zur vollständigen Beschreibung der Szenensegmentierung wird zusätzlich noch die Anzahl der aktuell im Bild vorhandenen, unabhängig bewegten Objekte ermittelt. Für die weitere Herleitung des Szenensegmentierung werden diese Größen jedoch zunächst

als bekannt vorausgesetzt. Ein probabilistisches Modell zur Szenensegmentierung wird aufgebaut, wobei die starken statistischen Bindungen in räumlicher und zeitlicher Richtung bei der Modellbildung explizit durch Markov-Zufallsfelder berücksichtigt werden. Die einzelnen Teile werden zu einem Gesamtmodell zusammengefasst, welches dann mit Hilfe aktueller Optimierungsverfahren gelöst werden kann.

KAPITEL 4

# Optimierungsstrategie und Parameterschätzung

Mit dem im letzten Kapitel aufgestellten Modell zur bewegungsbasierten Szenensegmentierung wird deutlich, dass die Güte der Segmentierung untrennbar mit der Qualität der Schätzung der Modellparameter verbunden ist. Die Aufgabe des Verfahrens besteht somit aus der gemeinsamen Lösung des in (3.2) beschriebenen Ausdrucks, der hier nochmals in Erinnerung gerufen werden soll:

$$\left\{\hat{\mathbf{l}}_t, \hat{\boldsymbol{\Theta}}_t\right\} = \arg\max_{\mathbf{l}_t, \boldsymbol{\Theta}_t} \left\{P(\mathbf{l}_t, \boldsymbol{\Theta}_t | \mathcal{G}_{t+1})\right\} .$$

Da eine gemeinsame Optimierung von Segmentierung und Modellparametern analytisch und rechentechnisch bereits für sehr einfache Aufgabenstellungen nicht mehr realisierbar ist, werden die einzelnen Ausdrücke getrennt voneinander optimiert[1]. In der Bildverarbeitung werden hierfür oftmals Relaxationsverfahren eingesetzt, die in regelmäßigen Abständen unterbrochen werden um dann für eine spezielle Realisation des Labelfeldes die Parameter neu zu schätzen. So wird z. B. in [Besag, 1986; Won u. Derin, 1992; Kato u. a., 1995; Li, 2009] durch deterministisches bzw. simuliertes Abkühlen das Labelfeld $\mathbf{l}_t$ optimiert, während darauf aufbauend eine Schätzung der Modellparameter durchgeführt wird. Die Zuweisung einer Beobachtung erfolgt hier eindeutig zu genau einer Objekthypothese. In der Literatur wird eine solche Abbildung von Beobachtungen auf eine Menge von Objekthypothesen als Datenassoziationsprozess bezeichnet. Das oben beschriebene Vorgehen stellt dabei eine „harte" Datenzuweisung (engl. *hard data assignment*) dar, d. h. bei der Schätzung der Parameter gehen die Beobachtungen vollständig in die Schätzung mit ein. Die Anzahl der Segmentklassen wird hierbei meist als bekannt vorausgesetzt.

[1] In der Literatur wird dies als *partial optimal solution* [Won u. Derin, 1992] bezeichnet. Ein bekanntes Problem bei einer solchen getrennten Optimierung ist die Gefahr, in lokalen Minima zu verharren.

Im Gegensatz dazu werden in [Bar-Shalom, 1987; Molnar u. Modestino, 1998; Hoffmann, 2007] Verfahren vorgestellt, bei denen die Datenzuweisung probabilistisch gewichtet efolgt. Das Anwendungsgebiet ist hier meist die Mehrzielverfolgung (engl. *multi-target tracking*) von punktförmig modellierbaren Objekten in unterschiedlichsten Umgebungen. Für eine gegebene Anzahl an Objekthypothesen wird hier die Schätzung des aktuellen Zustands durch eine „weiche" Datenassoziation (engl. *soft data asssignment*) verbessert. Der theoretisch durch den Assoziationsprozess entstehende hohe Rechenaufwand, der kombinatorisch mit der Anzahl an Objekthypothesen anwächst, kann dabei durch eine Überprüfung nur einer plausiblen Untermenge aller Assoziationsmöglichkeiten drastisch reduziert werden. In [Mahler u. Ronald, 2007] wird ein Verfahren vorgestellt, welches die Historie der Assoziationen zwischen Beobachtungen und zeitlich verfolgten Hypothesenparametern speichert und daraus die aktuelle Schätzung ableitet. Diese sog. Multi-Hypothesen-Trackingverfahren (MHT) (engl. *multiple hypothesis tracking*) skalieren jedoch mit der Anzahl der Hypothesen und sind dadurch sehr rechenaufwändig. Der von [Streit u. Luginbuhl, 1995] vorgestellte Ansatz kann als Erweiterung des MHT verstanden werden, wobei der Assoziationsprozess auf mehrerer Beobachtungen pro Hypothese erweitert wird. Die einzelnen Beobachtungen werden hier entsprechend ihrer Zuweisungswahrscheinlichkeit bzgl. einer Hypothese gewichtet und zu einer virtuellen Messung zusammengefasst. Diese gehen dann in die jeweilige Schätzung der Hypothesenparameter ein.

Alternativ dazu können die einzelnen Optimierungsschritte aber auch in verzahnter Reihenfolge abgearbeitet werden. Hierbei wird das Problem als Parameterschätzaufgabe mit „unvollständigen" Beobachtungen beschrieben. Die Beobachtungen werden hierbei als unvollständig in Bezug auf die Schätzung der Modellparameter bezeichnet [Dempster u. a., 1977; Gauvrit u. a., 1997]. Die Unvollständigkeit der Daten ergibt sich in der vorliegenden Arbeit aus der Tatsache, das sich das Bild aus einer unbekannten Menge sich unterschiedlich bewegender Objekte zusammensetzt und die Zugehörigkeit der einzelnen Bildpunkte zu diesen Objekten unbekannt ist. Die Objekthypothesen werden durch den Parametersatz $\boldsymbol{\theta}_t$ beschrieben. Ein häufig verwendetes Verfahren zur Lösung eines so formulierten Problems ist der EM-Algorithmus[2]. Hierbei handelt es sich um ein iteratives Verfahren, welches innerhalb eines Iterationsschrittes aus zwei, nacheinander ausgeführten, Operationen besteht: Ausgehend von der Schätzung der Parameter aus dem vorherigen Iterationsschritt, wird der Erwartungswert der Log-Likelihoodfunktion der vollständigen Daten ermittelt. Im darauf folgenden Schritt wird dann dieser Erwartungswert bzgl. der Parameter maximiert.

[2]Eine Einführung in das Problem der Parameterschätzung mit unvollständigen Daten und dessen Lösung mit dem EM-Verfahren wird in Anhang A.3 gegeben.

In dieser Arbeit wird die Idee des EM-Verfahrens aufgegriffen und auf das Problem der bewegungsbasierten Szenensegmentierung angepasst. Dabei werden die Beobachtungen nicht, entsprechend einer binären Entscheidungsregel, nur einer bestimmten Objekthypothese zugewiesen, sondern probabilistisch gewichtet. Entsprechend dem Modell des Labelprozesses aus Kapitel 3 wird eine Glattheitsanforderung an den Assoziationsprozesses in Form eines Markov-Zufallsfeldes beschrieben. Im Maximierungsschritt werden die Bewegungsparameter der einzelnen Objekthypothesen effizient durch ein rekursives Filter geschätzt und zeitlich nachgeführt.

## 4.1 Formale Beschreibung der Schätzaufgabe

Zur Beschreibung des Verfahrens werden die bereits bekannten Größen aus Kapitel 3 kurz in die gängige Taxonomie eingeordnet. Die in Abschnitt 3.3 vorgestellte Restfehlerkarte $\mathbf{E}_t$ stellt hierbei die unvollständigen Daten dar. Das Labelfeld mit den Realisationen $\mathbf{l}_t = (l_{1,t}, \ldots, l_{N,t})$ drückt die verdeckten, d. h. die unbekannten Daten aus. Der komplette Datensatz lautet somit $\tilde{\mathbf{E}}_t = \{\mathbf{E}_t, \mathbf{l}_t\}$ mit der Verbundverteilung $P(\tilde{\mathbf{E}}_t|\boldsymbol{\Theta}_t)$.

Als Erweiterung des herkömmlichen EM-Algorithmus wird a-priori Information bzgl. der Verteilung der Modellparameter mit eingebracht. Dabei wird angenommen, dass die zeitliche Entwicklung der Modellparameter einem Markov-Prozess entspringt und die Beobachtungen bei gegebener Segmentierung unabhängig von vorherigen Beobachtungen sind. Unter diesen Annahmen kann die Schätzaufgabe in rekursiver Form durch das bekannte Bayes-Filter ausgedrückt werden. Die Erweiterung des klassischen EM-Ansatzes um eine Komponente, die Vorwissen bzgl. der zu schätzenden Größe mit einbringt, wird in der Literatur auch als bestraftes (engl. *penalized*) EM-Verfahren bezeichnet [Green, 1990].

Das Verfahren beginnt damit, für eine gegebene Schätzung der Modellparameter $\hat{\boldsymbol{\Theta}}_t^k$, den Erwartungswert der vollständigen Daten zu berechnen. Aus Notationsgründen wird hierfür die Funktion

$$\mathrm{Q}(\boldsymbol{\Theta}_t|\hat{\boldsymbol{\Theta}}_t^k) = \mathrm{E}\left[\log P(\tilde{\mathbf{E}}_t|\boldsymbol{\Theta}_t, \mathbf{l}_{t-1})|\mathbf{E}_t, \hat{\boldsymbol{\Theta}}_t^k, \hat{\mathbf{l}}_{t-1}\right] \tag{4.1}$$

definiert. $k = (1, \ldots, K)$ bezeichnet dabei den Iterationsindex. Im zweiten Schritt innerhalb einer Iteration wird dann dieser Erwartungswert bzgl. der Modellparameter maximiert:

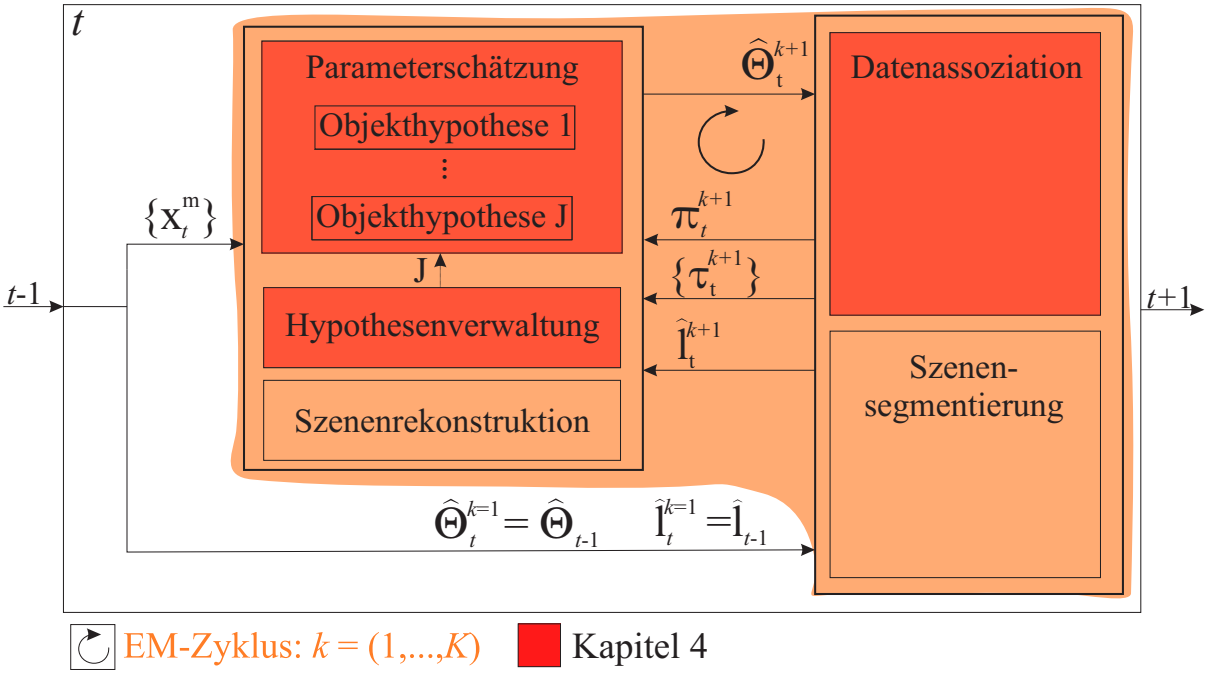



**Abbildung 4.1**: Ablaufdiagramm des iterativen Optimierungsverfahrens. Zu jedem Zeitpunkt $t$ werden die Größen $\hat{\mathbf{l}}_t^k$, $\hat{\mathbf{z}}_t^k$ und $\hat{\boldsymbol{\theta}}_t^k$ abwechselnd geschätzt. Die für einen Teil der Schätzung notwendigen Merkmalspunkte $\{\mathbf{x}_{\mathrm{n},t}^{\mathrm{m}} | \mathrm{n} \in \mathcal{K}\}$ werden dabei mit den jeweiligen Assoziationsgewichten $\boldsymbol{\tau}_{\mathrm{n},t}^k$ gewichtet.

$$\left\{\hat{\boldsymbol{\theta}}_t^{k+1}, \hat{\mathbf{z}}_t^{k+1}\right\} = \underset{\boldsymbol{\theta}_t, \mathbf{Z}_t}{\arg\max} \left\{ \mathrm{Q}(\boldsymbol{\Theta}_t | \hat{\boldsymbol{\Theta}}_t^k) + \log P(\boldsymbol{\Theta}_t | \mathbf{l}_t, \mathcal{G}_{t-1}) \right\} . \tag{4.2}$$

Durch die Berücksichtigung des zweiten Terms auf der rechten Seite von (4.2) ergibt sich eine zeitlich rekursive Formulierung der Schätzung bzgl. $\boldsymbol{\theta}_t$, wobei die prädizierte Schätzdichte zum Zeitpunkt $t-1$ als a-priori Wissen in die Schätzung mit eingeht. Zu jedem Zeitpunkt iteriert das Verfahren bis ein bestimmtes Abbruchkriterium erfüllt ist. Als Kriterien haben sich hier die relative Änderung $\|\hat{\boldsymbol{\theta}}_t^{k+1} - \hat{\boldsymbol{\theta}}_t^k\|$ des Parametervektors zwischen zwei Iterationsschritten und die Konvergenz des Erwartungwertes hin zu einem konstanten Wert bewährt.

Mit den so bestimmten Modellparametern, kann dann die Szenensegmentierung, wie in Kapitel 3 beschrieben, durchgeführt werden. Abbildung 4.1 verdeutlicht das Zusammenspiel der einzelnen Verarbeitungsschritte zu jedem Zeitpunkt $t$ des Schätzprozesses.

## 4.2 Bestimmung des Erwartungswertes

Die Bestimmung des Erwartungswertes der Log-Likelihoodfunktion gestaltet sich je nach Modellierung des Labelprozesses unterschiedlich aufwändig. Im Folgenden wird, ausgehend von sehr einfachen Annahmen, die Komplexität des Modells schrittweise erhöht und daraus eine mathematische Vorschrift zur Berechnung eines Erwartungswertes der gesuchten Segmentierung abgeleitet.

### 4.2.1 Modell statistischer Unabhängigkeit

Für den einfachen Fall unabhängiger Beobachtungen und zusätzlicher Unabhängigkeit des Labelprozesses, ergibt sich für die Log-Likelihood aus (4.1)

$$\begin{aligned}
\log P(\tilde{\mathbf{E}}_t|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) &= \log P(\mathbf{E}_t,\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) \\
&= \log P(\mathbf{E}_t|\mathbf{l}_t,\boldsymbol{\Theta}_t) + \log P(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) \\
&= \sum_{n\in\mathcal{R}} \log P(\boldsymbol{\varepsilon}_{n,t}|l_{n,t},\boldsymbol{\Theta}_t) + \sum_{n\in\mathcal{R}} \log P(l_{n,t}|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) \\
&= \sum_{n\in\mathcal{R}} \log P(\boldsymbol{\varepsilon}_{n,t},\mathrm{l}_{n,t}|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1})\,.
\end{aligned} \tag{4.3}$$

Um eine weiche Zuweisung der Daten in die bestehende Struktur des EM-Verfahrens integrieren zu können, ist es hilfreich, die Variablen $\mathbf{l}_t$ und $\boldsymbol{\Theta}_t$ voneinander zu trennen. Dies kann erreicht werden, indem anstatt der Zuweisung eines skalaren Wertes $j$ zu einer Komponente der Labelvariablen, ein $J$-dimensionaler Einheitsvektor verwendet wird, d. h. $\mathrm{l}_{n,t} = \mathbf{e}_j, j \in \mathcal{J} = \{1,\ldots,J\}$ für den Fall einer Zuweisung der Labelkomponente zu Objekthypothese $j$. Der Ausdruck in (4.3) kann dann umformuliert werden zu

$$\begin{aligned}
\log P(\tilde{\mathbf{E}}_t|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) &= \sum_{n\in\mathcal{R}} \sum_{j\in\mathcal{J}} \mathrm{l}_{n,t}^j \log P(\boldsymbol{\varepsilon}_{n,t},\mathrm{l}_{n,t} = \mathbf{e}_j|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) \\
&= \sum_{n\in\mathcal{R}} \mathrm{l}_{n,t}^{\mathrm{T}} \mathbf{U}(\boldsymbol{\varepsilon}_{n,t}|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1})\,, \text{wobei}
\end{aligned} \tag{4.4}$$

$\mathbf{U}(\boldsymbol{\varepsilon}_{n,t}|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) = [\log P(\boldsymbol{\varepsilon}_{n,t},\mathrm{l}_{n,t} = \mathbf{e}_1|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}),\ldots,\log P(\boldsymbol{\varepsilon}_{n,t},\mathrm{l}_{n,t} = \mathbf{e}_J|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1})]^{\mathrm{T}}$. Es gilt zu beachten, dass die einzelnen Komponenten hier konstant bzgl. $\mathrm{l}_{n,t}$ sind. Der Erwartungswert der Log-Likelihoodfunktion kann nun umgeschrieben werden

zu

$$\begin{aligned} \mathrm{Q}(\boldsymbol{\Theta}_t|\boldsymbol{\Theta}_t^k) &= \mathrm{E}\left[\sum_{n\in\mathcal{R}} \mathrm{l}_{n,t}^{\mathrm{T}}\mathbf{U}(\boldsymbol{\varepsilon}_{n,t}|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1})|\mathbf{E}_t,\hat{\boldsymbol{\Theta}}_t^k,\hat{\mathbf{l}}_{t-1}\right] \\ &= \sum_{n\in\mathcal{R}} \mathrm{E}\left[\mathrm{l}_{n,t}|\mathbf{E}_t,\hat{\boldsymbol{\Theta}}_t^k,\hat{\mathbf{l}}_{t-1}\right]\mathbf{U}(\boldsymbol{\varepsilon}_{n,t}|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}), \end{aligned} \tag{4.5}$$

mit dem Erwartungswert der verdeckten Daten

$$\mathrm{E}\left[\mathrm{l}_{n,t}|\mathbf{E}_t,\hat{\boldsymbol{\Theta}}_t^k,\hat{\mathbf{l}}_{t-1}\right] = \sum_{j\in\mathcal{J}} P(\mathrm{l}_{n,t}=\mathbf{e}_j|\boldsymbol{\varepsilon}_{n,t},\hat{\boldsymbol{\Theta}}_t^k,\hat{\mathbf{l}}_{t-1})\mathbf{e}_j = \sum_{j\in\mathcal{J}} \pi_{n,t}^j\mathbf{e}_j\,. \tag{4.6}$$

Eine Komponente dieser bedingten Wahrscheinlichkeit drückt das Gewicht einer Beobachtung an der Stelle $\mathbf{x}_{n,t}$ bzgl. der Objekthypothese $j$ aus und wird durch

$$\pi_{n,t}^j = \frac{P(\boldsymbol{\varepsilon}_{n,t}|\mathrm{l}_{n,t}=\mathbf{e}_j,\hat{\boldsymbol{\Theta}}_t^k)\Lambda_n^j}{\sum_{u\in\mathcal{J}} P(\boldsymbol{\varepsilon}_{n,t}|\mathrm{l}_{n,t}=\mathbf{e}_u,\hat{\boldsymbol{\Theta}}_t^k)\Lambda_n^u} \tag{4.7}$$

mit der Prioriverteilung

$$\Lambda_n^j = P(\mathrm{l}_{n,t}=\mathbf{e}_j|\hat{\boldsymbol{\Theta}}_t^k,\hat{\mathbf{l}}_{t-1}) \tag{4.8}$$

beschrieben. Der Erwartungswert in (4.1) kann dann folgendermaßen formuliert werden [3]

$$\mathrm{Q}(\boldsymbol{\Theta}_t|\boldsymbol{\Theta}_t^k) = \sum_{n\in\mathcal{R}} \boldsymbol{\pi}_{n,t}\left(\mathbf{D}(\boldsymbol{\varepsilon}_{n,t}|\boldsymbol{\Theta}_t)+\mathbf{V}(\boldsymbol{\Theta}_t,\mathbf{l}_{t-1})\right). \tag{4.9}$$

Die erste Komponente beschreibt hier die hypothesenspezifische Log-Likelihoodfunktion[4]

$$\begin{aligned} \log P(\boldsymbol{\varepsilon}_{n,t}|\mathrm{l}_{n,t},\boldsymbol{\Theta}_t) &= \mathrm{l}_{n,t}^{\mathrm{T}}\mathbf{D}(\boldsymbol{\varepsilon}_{n,t}|\boldsymbol{\Theta}_t) \quad \text{, mit} \\ \mathbf{D}(\boldsymbol{\varepsilon}_{n,t}|\boldsymbol{\Theta}_t) &= \left[\log P(\boldsymbol{\varepsilon}_{n,t}|\mathbf{e}_1,\boldsymbol{\Theta}_t),\ldots,\log P(\boldsymbol{\varepsilon}_{n,t}|\mathbf{e}_J,\boldsymbol{\Theta}_t)\right]^{\mathrm{T}}. \end{aligned} \tag{4.10}$$

[3] In ausführlicher Summenschreibweise:

$$\begin{aligned} \mathrm{Q}(\boldsymbol{\Theta}_t|\boldsymbol{\Theta}_t^k) &= \sum_{n\in\mathcal{R}}\sum_{j\in\mathcal{J}} \pi_{n,t}^j \log P(\boldsymbol{\varepsilon}_{n,t},\mathrm{l}_{n,t}=\mathbf{e}_j|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) \\ &= \sum_{n\in\mathcal{R}}\sum_{j\in\mathcal{J}} \pi_{n,t}^j \log P(\boldsymbol{\varepsilon}_{n,t}|\mathrm{l}_{n,t}=\mathbf{e}_j,\boldsymbol{\Theta}_t) + \sum_{n\in\mathcal{R}}\sum_{j\in\mathcal{J}} \pi_{n,t}^j \log P(\mathrm{l}_{n,t}=\mathbf{e}_j|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}). \end{aligned}$$

[4] Die Likelihood beschreibt die Wahrscheinlichkeit $P(\boldsymbol{\varepsilon}_{n,t}|\mathrm{l}_{n,t}=\mathbf{e}_j,\boldsymbol{\Theta}_t)=P(\boldsymbol{\varepsilon}_{n,t}|\mathbf{e}_j,\boldsymbol{\Theta}_t)$, mit welcher der Restfehler $\boldsymbol{\varepsilon}_{n,t}$ durch Objekthypothese $j$ erzeugt wurde.

In ähnlicher Weise kann die zweite Komponente umformuliert werden

$$\begin{aligned} \log P(\mathrm{l}_{n,t}|\boldsymbol{\Theta}_t, \mathbf{l}_{t-1}) &= \mathrm{l}_{n,t}^{\mathrm{T}} \mathbf{V}(\boldsymbol{\Theta}_t, \mathbf{l}_{t-1}), \text{ mit} \\ \mathbf{V}(\boldsymbol{\Theta}_t, \mathbf{l}_{t-1}) &= \left[\log P(\mathbf{e}_1|\boldsymbol{\Theta}_t, \mathbf{l}_{t-1}), \ldots, \log P(\mathbf{e}_J|\boldsymbol{\Theta}_t, \mathbf{l}_{t-1})\right]^{\mathrm{T}} . \end{aligned} \tag{4.11}$$

Bei der Auswertung des bedingten Erwartungwertes, kann durch $\Lambda^j$ Vorwissen[5] bzgl. des Zufallsprozesses eingebracht werden. Dieses Vorwissen kann aus objektspezifischen Eigenschaften bestehen, welche dem System vorab bekannt sind [Bachmann u. Balthasar, 2008; Bachmann u. Dang, 2008]. $\Lambda^j$ kann jedoch auch direkt aus den Daten abgeleitet werden, womit die Verteilung selbst eine Schätzgröße darstellt, die es zu bestimmen gilt.

Im Fall statistischer Unabhängigkeit des Labelprozesses, kann $\Lambda$ einfach durch Ableiten von (4.9) nach $\Lambda$ und anschließendem zu Null setzen ermittelt werden. Unter der Bedingung $\sum_{j\in\mathcal{J}} \Lambda^j = 1$ ergibt sich dann nach Lagrange folgender anschauliche Wert für die relative Häufigkeit, bzw. die mittlere Wahrscheinlichkeit der Objekthypothese:

$$\Lambda^j = \frac{1}{N} \sum_{n\in\mathcal{R}} \pi_{n,t}^j . \tag{4.12}$$

Wie bereits im letzten Kapitel erwähnt, führt die Annahme statistischer Unabhängigkeit durch Rauscheinflüsse und fehlerhafte Messungen zu einer unerwünscht hohen Fragmentierung bei der Zerlegung der Szene. Um solche Einflüsse zu reduzieren, wird ein Markov-Zufallsfeld eingeführt, mit welchem sich die Bedingung einer örtlichen und zeitlichen Glattheit der Segmentierung erzwingen lässt.

### 4.2.2 Modellierung statistischer Bindungen

Um die statistischen Bindungen des Assoziationsprozesses abzubilden, wird das bereits in Kapitel 3 eingeführte Modell eines Markov-Zufallsfeldes mit einem Nachbarschaftsystem erster Ordnung verwendet. Unter Ausnutzung der Gibbs-Markov-Äquivalenz ergibt sich dann für den zweiten Term auf der rechten Seite von (4.3) der Ausdruck

$$P(\mathbf{l}_t|\boldsymbol{\Theta}_t, \mathbf{l}_{t-1}) = \mathrm{Z}^{-1} \exp\left[-H(\mathbf{l}_t|\boldsymbol{\Theta}_t, \mathbf{l}_{t-1})\right] \tag{4.13}$$

[5] Ist keine Information über die Verteilung bekannt, handelt es sich bei $\Lambda^j$ um eine sog. nicht informative Prioriverteilung (engl. *non-informative prior*), was einer Gleichverteilung über alle Labelrealisationen entspricht.

mit dem Energiefunktional

$$H(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) = \sum_{n\in\mathcal{R}} \mathrm{l}_{n,t}^{\mathrm{T}}\mathbf{V}_1(\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) + \sum_{n,u\in\mathcal{C}} \mathrm{l}_{n,t}^{\mathrm{T}}\mathbf{V}_2(\boldsymbol{\Theta}_t)\mathrm{l}_{u,t}\ . \tag{4.14}$$

Die Vektorfunktion $\mathbf{V}_1(\boldsymbol{\Theta}_t,\mathbf{l}_{t-1})$ mit den Elementen $\log P(\mathrm{l}_{n,t}=\mathbf{e}_j|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1})$ wird hier durch das Modell in (3.18) beschrieben. $\mathbf{V}_2(\boldsymbol{\Theta}_t)$ kann als Matrix der Dimension $J\times J$ darstellt werden. $\log P(\mathrm{l}_{n,t}=\mathbf{e}_j,\mathrm{l}_{u,t}=\mathbf{e}_v|\boldsymbol{\Theta}_t)$ beschreibt dabei ein einzelnes Element dieser Matrix. Gleichung (3.16) liefert hier das Modell zur Bewertung der lokalen Nachbarschaft um jeden Punkt. Hieraus ergibt sich für den Erwartungswert des Labelprozesses[6]

$$\begin{aligned}\mathrm{Q}(\boldsymbol{\Theta}_t|\hat{\boldsymbol{\Theta}}_t^k) = \sum_{n\in\mathcal{R}} \boldsymbol{\pi}_{n,t}\left(\mathbf{D}(\varepsilon_{n,t}|\boldsymbol{\Theta}_t) - \mathbf{V}_1(\boldsymbol{\Theta}_t,\mathbf{l}_{t-1})\right) - \\ \sum_{n,u\in\mathcal{C}} \mathrm{E}\left[\mathrm{l}_{n,t}^{\mathrm{T}}\mathbf{V}_2(\boldsymbol{\Theta}_t)\mathrm{l}_{u,t}|\hat{\boldsymbol{\Theta}}_t^k\right] - \mathrm{E}\left[\mathrm{Z}|\hat{\boldsymbol{\Theta}}_t^k\right]\ .\end{aligned} \tag{4.15}$$

Die Hauptschwierigkeit bei der Auswertung dieses Ausdrucks hin zur Log-Likelihood der vollständigen Daten stellt die Bestimmung des bedingten Erwartungswertes in der zweiten Summe dar. So wird z. B. in [Zhang u. a., 1994] ein Monte-Carlo-Samplingverfahren verwendet, um aus der bedingten Verteilung der vollständigen Daten zufällig Stichproben zu ziehen und daraus den bedingten Erwartungswert zu berechnen. Aufgrund des relativ umfangreichen Stichprobenraums (ca. 300-1000 Samples pro Iterationsschritt) erweisen sich stochastische Verfahren in der Praxis aber als ineffizient und werden in dieser Arbeit nicht weiter untersucht. Stattdessen wird ein Näherungsverfahren verwendet, welches die wahre Verteilung der lokalen Charakteristik in (4.15) approximativ durch

$$\mathrm{E}\left[\mathrm{l}_{n,t}^{\mathrm{T}}\mathbf{V}_2(\boldsymbol{\Theta}_t)\mathrm{l}_{t,u}|\hat{\boldsymbol{\Theta}}_t^k\right] = \boldsymbol{\pi}_{n,t}^{k\,\mathrm{T}}\mathbf{V}_2(\boldsymbol{\Theta}_t)\boldsymbol{\pi}_{u,t}^k \tag{4.16}$$

[6] Gleichung (4.15) stellt eine Gibbs-Verteilung dar, wobei die Log-Likelihood als Cliquenfunktion einer Einerclique betrachtet werden kann

$$\begin{aligned}P(\tilde{\mathbf{E}}_t|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) &= \prod_{n\in\mathcal{R}} P(\varepsilon_{n,t}|\mathrm{l}_{n,t},\hat{\boldsymbol{\Theta}}_t^k)\ \mathrm{Z}^{-1}\exp\left[-H\left(\mathbf{l}_t|\hat{\boldsymbol{\Theta}}_t^k,\hat{\mathbf{l}}_{t-1}\right)\right] \\ &= \exp\left[\sum_{n\in\mathcal{R}} \log P(\varepsilon_{n,t}|\mathrm{l}_{n,t},\hat{\boldsymbol{\Theta}}_t^k) - H\left(\mathbf{l}_t|\hat{\boldsymbol{\Theta}}_t^k,\hat{\mathbf{l}}_{t-1}\right)\right]\mathrm{Z}^{-1}\end{aligned}$$

beschreibt. Die Bestimmung des Erwartungswerts wird hier auf der Basis der bekannten Pseudo-Likelihood [Besag, 1986] durchgeführt. Hierbei werden die statistischen Abhängigkeiten auf das lokale Nachbarschaftssystem $\mathcal{N}$ begrenzt in der Form

$$P(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) = \prod_{n\in\mathcal{R}} P(\mathrm{l}_{n,t}|\mathrm{l}_{u,t}, u\in\mathcal{N}_n, \boldsymbol{\Theta}_t, \mathbf{l}_{t-1}) , \tag{4.17}$$

woraus folgt

$$\begin{aligned} P(\mathrm{l}_{n,t}|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) &= \sum_{\substack{\mathrm{l}_{u,t}\\ u\neq n}} P(\mathbf{l}_t|\boldsymbol{\Theta}_t,\mathbf{l}_{t-1}) \\ &\approx \sum_{\substack{\mathrm{l}_{u,t}\\ u\neq n}} \prod_{u\in\mathcal{R}} P(\mathrm{l}_{u,t}|\mathrm{l}_{w,t}, w\in\mathcal{N}_u, \boldsymbol{\Theta}_t, \mathbf{l}_{t-1}) . \end{aligned} \tag{4.18}$$

Mit dem Schätzergebnis $\boldsymbol{\pi}_{w,t}^{k-1}$ aus dem letzten Iterationsschritt ergibt sich die Prioriverteilung einer einzelnen Komponente in (4.7) dann näherungsweise durch

$$\begin{aligned} \Lambda_{n,t}^{j} &= P(\mathrm{l}_{n,t}=\mathbf{e}_j|\boldsymbol{\pi}_{u,t}, u\in\mathcal{N}_n, \hat{\boldsymbol{\Theta}}_t^k, \hat{\mathbf{l}}_{t-1}) \\ &\approx \frac{\exp\left[-H\left(\mathrm{l}_{n,t}=\mathbf{e}_j|\boldsymbol{\pi}_{u,t}^{k-1}, u\in\mathcal{N}_n, \hat{\boldsymbol{\Theta}}_t^k, \hat{\mathbf{l}}_{t-1}\right)\right]}{\sum_{v\in\mathcal{J}}\exp\left[-H\left(\mathrm{l}_{n,t}=\mathbf{e}_v|\boldsymbol{\pi}_{u,t}^{k-1}, u\in\mathcal{N}_n, \hat{\boldsymbol{\Theta}}_t^k, \hat{\mathbf{l}}_{t-1}\right)\right]} \end{aligned} \tag{4.19}$$

berechnet werden. Die Näherung in obiger Gleichung ist dem ICM-Verfahren (siehe Anhang A.1) ähnlich. Es unterscheidet sich jedoch durch die Verwendung „weicher" Werte für $\mathrm{l}_{u,t}$, womit diese Art der Bestimmung der bedingten Verbundwahrscheinlichkeit dem sog. Mean-Field-Verfahren von [Zhang, 1992] sehr ähnlich ist. Das Konvergenzverhalten bei einer weichen Zuweisung der Daten ist im Vergleich zu einer harten Zuweisung, wie z. B. beim ICM-Verfahren, deutlich besser. Jedoch muss erwähnt werden, dass wie bei allen EM-Verfahren, Konvergenz nur bzgl. eines lokalen Minimums garantiert werden kann.

## 4.3 Schätzung der Parameter des Objektmodells

Die Schätzung der Modellparameter $\boldsymbol{\theta}_t \subset \boldsymbol{\Theta}_t$ des Objektmodells vereinfacht sich durch die Annahme gegenseitiger Unabhängigkeit der einzelnen Hypothesenparameter. Hierdurch kann die Optimierungsaufgabe in $J$ parallele Maximierungen der Q-Funktion zerlegt werden. Im Folgenden wird mit einer bildweisen Auswertung

der Daten begonnen, die zu jedem Zeitpunkt die einzelnen Modellparameter auf der Basis der zur Verfügung stehenden Beobachtungen bestimmt. Im Anschluss daran wird die Beschreibung der Schätzaufgabe in zeitlicher Richtung erweitert, womit auch Information über die Systemdynamik in das Schätzergebnis mit einfließt.

### 4.3.1 Bildweise Auswertung der Daten

Bei der bildweisen Auswertung wird zu jedem Zeitpunkt $t$ der Ausdruck

$$\sum_{n\in\mathcal{R}} \boldsymbol{\pi}_{n,t}^{k} \mathrm{D}(\varepsilon_{n,t}|\boldsymbol{\theta}_t)\,, \text{ bzw. } \sum_{n\in\mathcal{R}}\sum_{j\in\mathcal{J}} \pi_{n,t}^{j,k} \log P(\varepsilon_{n,t}|\mathrm{l}_{n,t}=\mathbf{e}_j,\boldsymbol{\theta}_t) \tag{4.20}$$

bzgl. $\boldsymbol{\theta}_t$ maximiert. Für die weiteren Herleitungen wird, stellvertretend für alle Objekthypothesen, die Schätzung der Modellparameter $\boldsymbol{\theta}_t^j$ für eine einzelne Objekthypothese $j$ vorgestellt. (4.20) vereinfacht sich dann zu

$$\hat{\boldsymbol{\theta}}_t^{\,j} = \arg\max_{\boldsymbol{\theta}_t^j} \left\{ \sum_{n\in\mathcal{R}} \pi_{n,t}^{j,k} \log P(\varepsilon_{n,t}|\mathrm{l}_{n,t}^{j},\boldsymbol{\theta}_t^j) \right\} \quad \forall\, j=1,\ldots,J\,. \tag{4.21}$$

Auf die Verwendung von Index $j$ wird im Folgenden, soweit dies ohne Verlust an Klarheit möglich ist, zugunsten einer kompakten Darstellung verzichtet. Mit dem in Abschnitt 3.3 eingeführten Modell einer normalverteilten Likelihoodfunktion, ergeben sich aus obiger Linearkombination die Momente der Likelihoodverteilung somit als gewichtetes Mittel der Messungen. Für die Parameter der Restfehlerverteilung gilt dann

$$\boldsymbol{\phi}_t^j = \frac{\sum_{n\in\mathcal{R}} \pi_{n,t}^{j} \varepsilon_{n,t}^2}{\sum_{n\in\mathcal{R}} \pi_{n,t}^{j}}\,. \tag{4.22}$$

Der Nenner in obiger Gleichung kann als effektive Anzahl der Bildpunkte interpretiert werden, die der Objekthypothese $j$ zugewiesen werden. Die Schätzung der Lageparameter $\boldsymbol{\xi}_t = (\mathbf{M}_t, \boldsymbol{\Sigma}_t)$ erfolgt auf sehr ähnlich Weise, jedoch nicht auf Basis der Restfehlerkarte selbst. Durch die Auswertung der Tiefenkarte $\mathbf{z}_t$ ergibt sich die Objektlage im Raum

$$\mathbf{M}_t^j = \frac{\sum_{n\in\mathcal{R}} \pi_{n,t}^{j} \mathbf{X}_{n,t}}{\sum_{n\in\mathcal{R}} \pi_{n,t}^{j}} \tag{4.23}$$

als das gewichtete Mittel der rekonstruierten Szenenpunkte $\mathbf{X}_{n,t}$. Für die Objektausdehnung folgt

$$\mathbf{\Sigma}_t^j = \frac{\sum_{n\in\mathcal{R}} \pi_{n,t}^j (\mathbf{X}_{n,t} - \mathbf{M}_t)^2}{\sum_{n\in\mathcal{R}} \pi_{n,t}^j} \,. \tag{4.24}$$

### Merkmalsbasierte Schätzung der Bewegungsparameter

Die Bestimmung der Bewegungsparameter erfolgt über ein indirektes Verfahren. Solche Verfahren zeichnen sich durch die Verwendung einer Menge korrespondierender Merkmalspunkte $\{\mathbf{x}_\mathrm{n}^\mathrm{m} | \mathrm{n} \in \mathcal{K}\}$ aus, wie in Abschnitt 2.2 im Kontext der Bewegtbildanalyse bereits erläutert wurde. Die in dieser Arbeit verwendeten Merkmale drücken dabei die Korrespondenzen im linken und nächsten rechten Kamerabild bzgl. der Merkmalspunkte in der rechten Referenzansicht aus. Die Assoziationswahrscheinlichkeiten der Daten aus dem letzten Abschnitt werden entsprechend

$$\boldsymbol{\tau}_{\mathrm{n},t}^j = \frac{\pi_{\mathrm{n},t}^j}{\sum_{u\in\mathcal{J}} \pi_{\mathrm{n},t}^u} \quad \forall\, \mathrm{n} \in \mathcal{K} = \{1,\ldots,M\} \tag{4.25}$$

auf die Merkmalspunkte abgebildet. Jede Komponente $\boldsymbol{\tau}_{\mathrm{n},t} = (\tau_{\mathrm{n},t}^1, \ldots, \tau_{\mathrm{n},t}^J)^\mathrm{T}$ drückt die jeweilige Wahrscheinlichkeit einer Objekthypothese an Position $\mathbf{x}_\mathrm{n}$ im Bild aus. Es gilt die Bedingung $\sum_{j\in\mathcal{J}} \tau_{\mathrm{n},t}^j = 1$.

Als Beobachtung werden die Verschiebungsvektoren $\{\mathbf{d}_{\mathrm{n},t} | \mathrm{n} \in \mathcal{K}\}$ bestimmt und für die flussbasierte Bewegungsschätzung verwendet. Zusätzlich werden die Tiefenwerte $\boldsymbol{\rho}_t = (\rho_{1,t}, \ldots, \rho_{M,t})$ der Merkmale bestimmt. Zur Verbesserung der Schätzergebnisse werden die Merkmalspunkte zeitlich verfolgt. Hierfür wurde neben dem KLT-Tracker [Shi u. Tomasi, 1994] auch ein korrelationsbasiertes Verfahren auf der Basis des Harris- [Harris u. Stephens, 1988] und des FAST-Eckendetektors [Rosten u. Drummond, 2006], sowie verteilungsbasierte Verfahren wie der SIFT-[Lowe, 2004] und der SURF-Deskriptor [Bay u. a., 2008] verwendet. Abbildung 4.2 zeigt die Merkmalspunkte mit dem gemessenen Verschiebungsvektorfeld $\{\mathbf{d}_{\mathrm{n},t} | \mathrm{n} \in \mathcal{K}\}$ in blau. Diese Messungen werden zur Schätzung der Bewegung genutzt. In rot und grün ist das daraus abgeleitete, erwartete Verschiebungsfeld der jeweiligen Szenensegmente eingeblendet.

Das Beobachtungsmodell des Schätzers ist definiert durch

$$\varsigma_{\mathrm{n},t} = \mathbf{h}(\boldsymbol{\theta}_t, \mathbf{x}_{\mathrm{n},t}^\mathrm{m}) + \mathbf{e}_t \,, \tag{4.26}$$

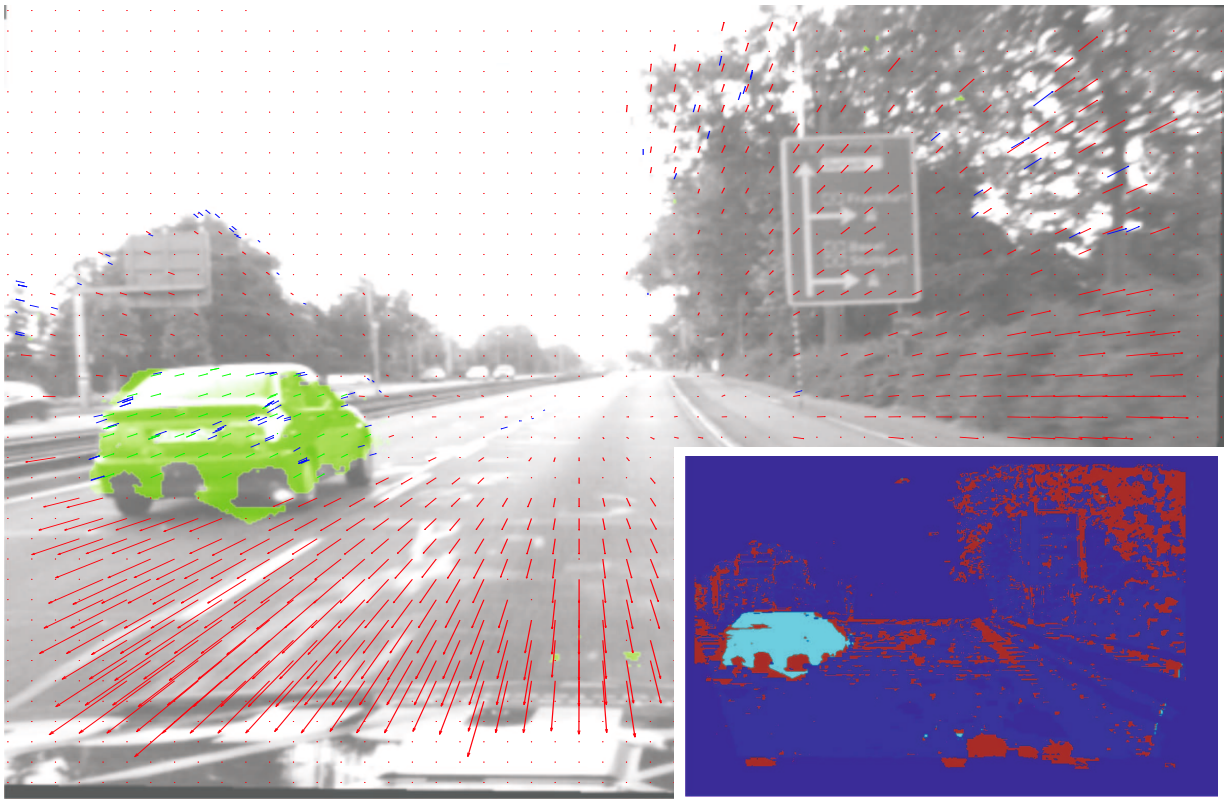


**Abbildung 4.2**: Verschiebungsvektoren der Merkmalspunkte in blau und das aus den entsprechenden Bewegungsschätzungen abgeleitete, dichte Verschiebungsfeld der Szene in rot und grün. Aus Darstellungsgründen ist nur eine Untermenge dieses Feldes visualisiert. Unten rechts ist die dazugehörige Segmentierung zu sehen.

wobei das Beobachtungsrauschen $\mathbf{e}_t$ als weiß, gaussverteilt mit Kovarianzmatrix $\Sigma_{\mathbf{ee},t} = \mathrm{E}\left[\mathbf{e}_t \mathbf{e}_t^{\mathrm{T}}\right]$ modelliert wird. Eine einzelne Komponente der Likelihoodfunktion wird demnach durch

$$P(\varsigma_{\mathrm{n},t} | \mathrm{l}^j_{\mathrm{n},t}, \boldsymbol{\theta}_t)^{\tau^j_{\mathrm{n},t}} = \mathcal{N}(\mathbf{h}(\boldsymbol{\theta}_t, \mathbf{x}^{\mathrm{m}}_{\mathrm{n},t}), \Sigma^*_{\mathbf{ee},\mathrm{n},t}) \tag{4.27}$$

beschrieben, wobei das Assoziationsgewicht in die Kovarianzmatrix der Messung integriert wird [Koch u. Yang, 1998; Younis, 1996; Ting u. a., 2007], d. h. $\Sigma^*_{\mathbf{ee},\mathrm{n},t} = (\tau^j_{\mathrm{n},t})^{-1} \Sigma_{\mathbf{ee},\mathrm{n},t}$. In Abbildung 4.3 ist das jeweilige Assoziationsgewicht $\tau^j_{\mathrm{n},t}$ einer Objekthypothese zusammen mit dem entsprechenden Residuum $r_{\mathrm{n},t} = |(\boldsymbol{\varsigma}_{\mathrm{n},t} - \hat{\boldsymbol{\varsigma}}_{\mathrm{n},t})^{\mathrm{T}} (\boldsymbol{\varsigma}_{\mathrm{n},t} - \hat{\boldsymbol{\varsigma}}_{\mathrm{n},t})|$ für eine Menge von Merkmalspunkten gezeigt. Das obere Bild zeigt die Verteilung der Gewichte und Residuen kurz nach der Initialisierung der Bewegungsschätzung. Das untere Diagramm zeigt die Verteilungen nach mehreren Iterationen. Zur Bewertung der Assoziationsgewichte wurde die mittlere Gewichtung der Punkte, die oberhalb des Medianwertes aller Residuen $r_{\mathrm{med}}$ lag, mit dem mittleren Gewicht der Punkte unterhalb dieses Wertes

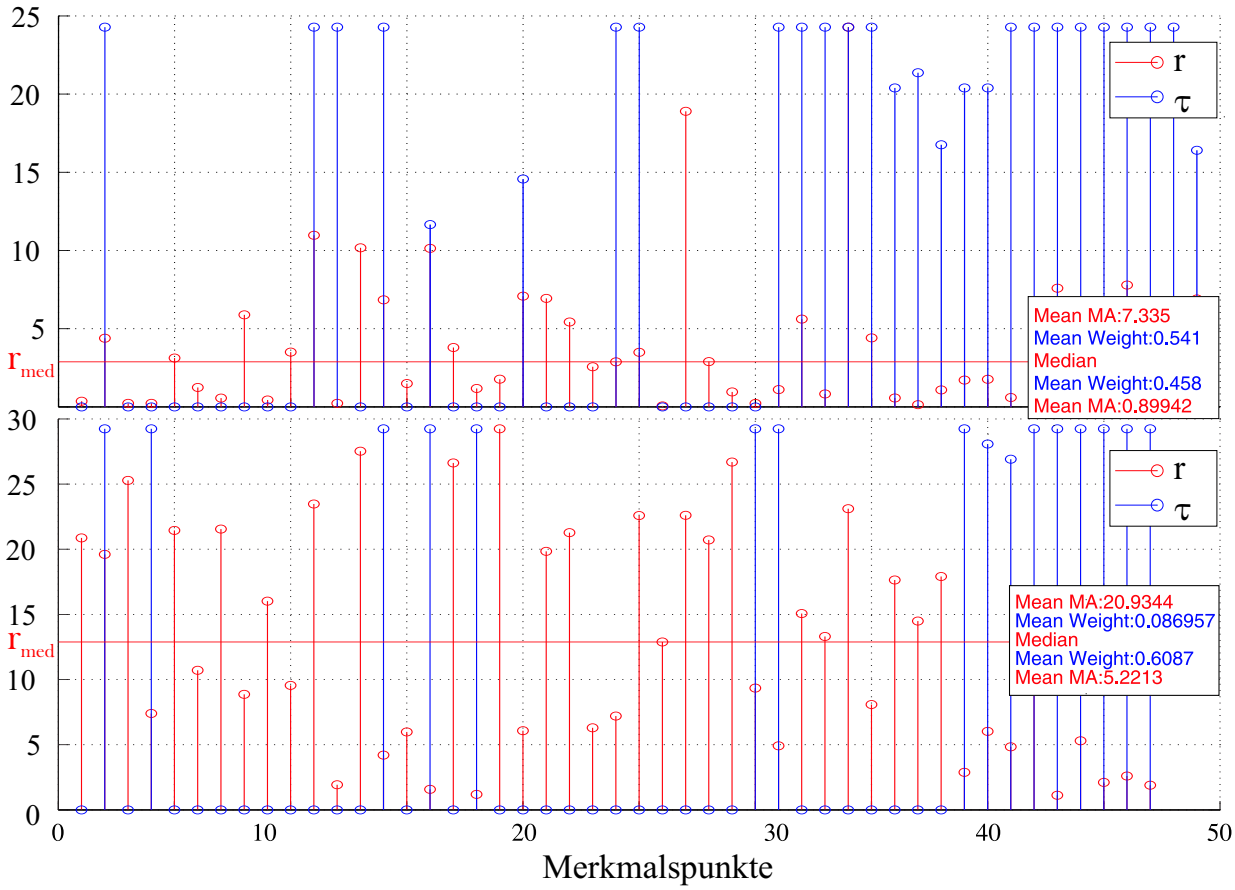



**Abbildung 4.3**: Die Residuenbeträge $r_{\mathrm{n},t}$ in rot und die entsprechenden Assoziationsgewichte $\tau_{\mathrm{n},t}$ in blau für eine Menge von Merkmalspunkten. Der Medianwert des Residuums $r_{\mathrm{med}}$ ist durch eine rote Linie markiert. Oben: Die Verteilung der Werte kurz nach der Initialisierung. Unten: Die entsprechende Verteilung nach Konvergenz der Bewegungsschätzung. Die anfangs noch annähernd gleich verteilten Gewichte sinken durch die verbesserte Bewegungsschätzung für Punkte mit kleinen Residuen, während Punkte mit hohen Residuen kleiner gewichtet werden.

verglichen. Bei der Auswertung anhand verschiedener Verkehrssituationen und Objektkonstellationen konnte ein deutlicher Zusammenhang zwischen den beiden Größen festgestellt werden. Der Anstieg von $r$ im unteren Diagramm von Abbildung 4.3 ist dadurch zu erklären, dass die Schätzung alle Merkmalspunkte mit einbezieht und entsprechend ihrer Assoziationswahrscheinlichkeit gewichtet. Punkte, die auf anders bewegten Objekten liegen, erzeugen somit hohe Residuenbeträge, welche bei klassischen Ansätzen als Ausreißer eingestuft und verworfen werden.

Als zu minimierende Gütefunktion bzgl. der gemessenen Daten ergibt sich[7]

[7]Im weiteren Verlauf wird für das Distanzmaß innerhalb des Klammerausdrucks auf der rechten Seite auch die abkürzende Schreibweise $\|\varsigma_{\mathrm{n},t} - \hat{\varsigma}_{\mathrm{n},t}\|^2_{\Sigma^*_{\mathbf{ee},\mathrm{n},t}}$ verwendet.

$$C_{\mathrm{m}} = \sum_{\mathrm{n}\in\mathcal{K}} \left(\boldsymbol{\varsigma}_{\mathrm{n},t} - \hat{\boldsymbol{\varsigma}}_{\mathrm{n},t}\right)^{\mathrm{T}} \left(\Sigma^{*}_{\mathbf{ee},\mathrm{n},t}\right)^{-1} \left(\boldsymbol{\varsigma}_{\mathrm{n},t} - \hat{\boldsymbol{\varsigma}}_{\mathrm{n},t}\right) . \tag{4.28}$$

Dieses mit der inversen Kovarianz gewichtete Maß wird auch als Mahalanobisnorm [Mahalanobis, 1936] bezeichnet. Die Minimierung des Gütemaßes in (4.28) definiert den Maximum-Likelihood-Schätzer[8], der bei einem gegebenen Datensatz für zeitlich konstante Parameter eine optimale Schätzung ergibt.

Wie bei anderen bekannten Verfahren zur Datenassoziation, stellt die Komplexität des auszuwertenden Ausdrucks auch hier eine große Hürde für die praktische Anwendung dar. Eine gebräuchliche Methode zur Beschränkung des Rechenaufwandes ist in der Literatur unter dem Begriff „Gating" bekannt. Hierbei wird der Messraum auf eine, für die jeweilige Hypothese plausible, Untermenge reduziert. Bei dem hier vorgestellten Verfahren wird prinzipiell auf einen solchen, meist heuristisch motivierten, Gatingprozess verzichtet. Jedoch enthält die formale Beschreibung bereits eine Art Gatingprozess wie aus (4.28) ersichtlich ist: setzt man $\boldsymbol{\tau}^{j}_{\mathrm{n},t} = 0$ trägt die jeweilige Beobachtung mit dem multiplikativen Faktor $+1$ zur Schätzung bei. Somit haben mit „0" gewichtete Beobachtungen keinen Einfluss auf die Schätzung. Die Verwendung der „harten" Datenzuweisung ($\tau^{j}_{\mathrm{n},t} \rightarrow \hat{\mathrm{l}}^{j}(\mathbf{x}_{\mathrm{n},t})$) aus dem Segmentierprozess kann hier als Spezialfall betrachtet werden, der alle Beobachtungen mit $\hat{\mathrm{l}}^{j}(\mathbf{x}_{\mathrm{n},t}) = \mathbf{e}_{j}$ gleichermaßen für die Schätzung berücksichtigt. Bekannte Vertreter dieser Art der Datenzuweisung sind das *k-means*-Verfahren und der *mean-shift*-Algorithmus. In [Bishop, 2006] findet sich hierzu eine ausführliche Beschreibung.

Für eine bildweise Bestimmung der Bewegungsparameter, wie z. B. in [Horn u. a., 2007] beschrieben, wird das in Abschnitt 2.2 vorgestellte Bewegungsmodell starrer Körper verwendet. Als Beobachtungen für die flussbasierte Bewegungsschätzung werden hier die Verschiebungsvektoren $\{\mathbf{d}_{\mathrm{n},t} | \mathrm{n} \in \mathcal{K}\}$ der extrahierten und zeitlich verfolgten Merkmalspunkte verwendet. Die zusätzlich benötigen Tiefenwerte $\boldsymbol{\rho}_{t}$ der Punkte werden in dem Modell als konstant angenommen.

[8] Die Likelihooddichte $P(\mathbf{y}|\boldsymbol{\theta})$ drückt die Dichte der Beobachtungen $\mathbf{y}$ bei bekanntem Parametervektor $\boldsymbol{\theta}$ aus. Im Gegensatz dazu drückt die a-posteriori Dichte $P(\boldsymbol{\theta}|\mathbf{y})$ die Dichte des Parametervektors bei bekannten Beobachtungen aus. Wünschenswert ist ein Schätzer $\hat{\boldsymbol{\theta}}$ , an der die a-posteriori Dichte ein Maximum hat. Beim sog. Gauss-Markov-Schätzer wird die Annahme getroffen, das neben der Likelihooddichte $\mathcal{N}(\mathbf{h}(\boldsymbol{\theta}_{t}, \mathbf{x}^{\mathrm{m}}_{t}), \Sigma_{\mathbf{ee}})$ auch der Parametervektor $\mathcal{N}(\boldsymbol{\theta}, \Sigma_{\mathbf{qq}})$ normalverteilt ist. Die Schätzung ergibt dann $\hat{\boldsymbol{\theta}} = (\mathbf{C}^{\mathrm{T}}\Sigma^{-1}_{\mathbf{ee}}\mathbf{C} + \Sigma^{-1}_{\mathbf{qq}})^{-1}\mathbf{C}^{\mathrm{T}}\Sigma^{-1}_{\mathbf{ee}}\mathbf{y}$, wobei $\mathbf{C}$ das Messmodell beschreibt. Ist die Statistik des Zustandsvektors unbekannt, also z. B. eine Normalverteilung mit unendlicher Varianz $\Sigma^{-1}_{\mathbf{qq}} = 0$, reduziert sich die Schätzung auf den oben erwähnten ML-Schätzer $\hat{\boldsymbol{\theta}} = (\mathbf{C}^{\mathrm{T}}\Sigma^{-1}_{\mathbf{ee}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\Sigma^{-1}_{\mathbf{ee}}\mathbf{y}$. Wird außerdem noch die Annahme weißen Beobachtungsrauschens $\Sigma_{\mathbf{ee}} = \sigma\mathbf{I}$ getroffen, ergibt sich der LS-Schätzer $\hat{\boldsymbol{\theta}} = (\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{y}$.

Für eine Anzahl von $M$ Merkmalspunkten ergibt sich dann die Beobachtungsgleichung

$$\underbrace{\begin{pmatrix} \mathbf{d}_{1,t} \\ \vdots \\ \mathbf{d}_{\mathrm{n},t} \\ \vdots \\ \mathbf{d}_{M,t} \end{pmatrix}}_{=:\hat{\boldsymbol{\varsigma}}_t} = \underbrace{\begin{pmatrix} \mathbf{C}_{1,t} \\ \vdots \\ \mathbf{C}_{\mathrm{n},t} \\ \vdots \\ \mathbf{C}_{M,t} \end{pmatrix}}_{=:\mathbf{C}_t} \begin{pmatrix} \boldsymbol{\omega}_t \\ \mathbf{t}_t \end{pmatrix} + \mathbf{e}_t = \mathbf{C}_t \mathbf{v}_t^{\mathrm{T}} + \mathbf{e}_t \,. \tag{4.29}$$

Die Matrix $\mathbf{C}_t$ enthält hier, neben den Parametern des Bewegungsmodells wie in (2.9) beschrieben, zusätzlich den als konstant angenommenen Tiefenwert für jeden Merkmalspunkt. Die Bewegungsparameter können dann einfach durch Lösen der Gleichung

$$\hat{\mathbf{v}}_t^{\mathrm{T}} = \left(\mathbf{C}^{\mathrm{T}} \Sigma_{\mathbf{ee},t}^{*-1} \mathbf{C}\right)^{-1} \mathbf{C}^{\mathrm{T}} \Sigma_{\mathbf{ee},t}^{*-1} \hat{\boldsymbol{\varsigma}}_t \tag{4.30}$$

bestimmt werden. Der Vorteil einer weichen Datenzuweisung für die Schätzung der Modellparameter macht sich vor allem bei der Neuinitialisierung von Objekthypothesen deutlich bemerkbar. Die anfangs noch stark unsicherheitsbehafteten Schätzergebnisse führen zu meist fehlerhaften Segmentierergebnissen, die bei einer harten Zuweisung der Daten oftmals zu einer unmittelbaren Deinitialisierung der entsprechenden Hypothese führen. Mit der weichen Datenzuweisung kann dieser Effekt merklich reduziert werden, womit potentielle Objekte über die kritische „Startphase" hinweg gerettet werden können.

### 4.3.2 Sequentielle Auswertung der Daten

Die bisherige Betrachtung hat sich auf die Auswertung von Beobachtungen zu einem Zeitpunkt beschränkt. Liegen sequentielle Daten vor, ist es jedoch wünschenswert, die Daten der gesamtem Messreihe auszuwerten. Dies gilt speziell für die zu bestimmenden Bewegungsparameter. Betrachtet man die gesamte Stereobildsequenz der Länge $\mathcal{T} = \{1, \ldots, T\}$, ist die Log-Likelihood der vollständigen Daten gegeben durch [Gauvrit u. a., 1997]

$$\begin{aligned} \mathrm{Q}(\boldsymbol{\Theta}_{\mathcal{T}} | \boldsymbol{\Theta}_{\mathcal{T}}^k) = & \sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{R}} \sum_{j_{n,t} \in \mathcal{J}_t} \pi_{n,t}^{j_{n,t}} \log P(\boldsymbol{\varepsilon}_{n,t} | \mathrm{l}_{n,t} = \mathbf{e}_j, \boldsymbol{\Theta}_t) \\ & + \sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{R}} \sum_{j_{n,t} \in \mathcal{J}_t} \pi_{n,t}^{j_{n,t}} \log P(\mathrm{l}_{n,t} = \mathbf{e}_j | \boldsymbol{\Theta}_t) \,. \end{aligned} \tag{4.31}$$

Unter der Annahme einer konstanten Anzahl von Objekthypothesen innerhalb der Messreihe, ergibt sich durch einfache Umsortierung hieraus

$$\begin{aligned} \mathrm{Q}(\boldsymbol{\Theta}_{\mathcal{T}}|\boldsymbol{\Theta}_{\mathcal{T}}^{k}) = \sum_{j_n\in\mathcal{J}}\sum_{t\in\mathcal{T}}\left[\sum_{n\in\mathcal{R}}\pi_{n,t}^{j}\right]\log P(\mathrm{l}_{n,t}^{j}|\boldsymbol{\Theta}_t) \\ + \sum_{j_n\in\mathcal{J}}\sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{R}}\pi_{n,t}^{j}\log P(\boldsymbol{\varepsilon}_{n,t}|\mathrm{l}_{n,t}=\mathbf{e}_j,\boldsymbol{\Theta}_t)\,. \end{aligned} \tag{4.32}$$

Hierbei wird angenommen, dass die einzelnen Komponenten der Assoziationsvariable $\mathbf{l}_t$ statistisch unabhängig sind. In einem solchen Fall kann die Maximierung der Modellparameter durch Erweiterung von (4.21) über alle Zeitschritte folgendermaßen geschrieben werden

$$\hat{\boldsymbol{\theta}}_{T}^{j} = \arg\max_{\boldsymbol{\theta}_T^j}\left\{\sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{R}}\pi_{n,t}^{j,k}\log P(\boldsymbol{\varepsilon}_{n,t}|\mathrm{l}_{n,t}^{j},\boldsymbol{\theta}_t^{j})\right\}. \tag{4.33}$$

Weiterhin ist es wünschenswert, die zeitliche Entwicklung der Systemparameter mit zu berücksichtigen. Hierbei angenommen, dass sich die Hypothesenparamter als Markov-Prozess erster Ordnung beschreiben lassen, d. h.

$$\begin{aligned} \log P(\boldsymbol{\theta}_{\mathcal{T}}) &= \log\left(P(\boldsymbol{\theta}_0)\prod_{\substack{t\in\mathcal{T}\\ t\neq 1}}P(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\right) \\ &= \log\left(\prod_{j\in\mathcal{J}}P(\boldsymbol{\theta}_0^{j})\prod_{j\in\mathcal{J}}\prod_{\substack{t\in\mathcal{T}\\ t\neq 1}}P(\boldsymbol{\theta}_t^{j}|\boldsymbol{\theta}_{t-1}^{j})\right) \\ &= \sum_{j\in\mathcal{J}}\log P(\boldsymbol{\theta}_0^{j}) + \sum_{j\in\mathcal{J}}\sum_{\substack{t\in\mathcal{T}\\ t\neq 1}}\log P(\boldsymbol{\theta}_t^{j}|\boldsymbol{\theta}_{t-1}^{j})\,. \end{aligned} \tag{4.34}$$

Der letzte Term beschreibt hier die zeitliche Entwicklung des Systemzustandes. Für den Fall eines linearen Systemodells mit Gauss'schem Rauschanteil gilt

$$\boldsymbol{\theta}_{t+1} = \mathbf{f}(\boldsymbol{\theta}_t) + \mathbf{q}_t;\ \ \mathbf{q}_t \sim \mathcal{N}(0,\Sigma_{\mathbf{qq},t})\,. \tag{4.35}$$

Hierbei beschreibt $\mathbf{f}(\boldsymbol{\theta}_t)$ die Dynamik des Systems. Additive Störungen $\mathbf{q}$ werden dabei als mittelwertfreie Gauss'sche Zufallsgröße mit Kovarianz $\Sigma_{\mathbf{qq},t} = \mathrm{E}\left[\mathbf{q}_t\mathbf{q}_t^{\mathrm{T}}\right]$ modelliert. Diese wird oftmals auch als System- bzw. Prozessrauschen bezeichnet.

Für eine sequentielle Auswertung der Daten ergibt sich mit dem oben vorgestellten Modell dann folgendes zu minimierendes Gütekriterium:

$$C_{\mathrm{m}} = \|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_0\|^2_{\boldsymbol{\Sigma}_{\mathbf{qq},0}} + \sum_{t\in\mathcal{T}} \|\boldsymbol{\theta}_t - \mathbf{f}(\hat{\boldsymbol{\theta}}_{t-1})\|^2_{\boldsymbol{\Sigma}_{\mathbf{qq},t}} + \sum_{\mathrm{n}\in\mathcal{K}} \|\boldsymbol{\varsigma}_{\mathrm{n},t} - \hat{\boldsymbol{\varsigma}}_{\mathrm{n},t}\|^2_{\boldsymbol{\Sigma}^*_{\mathbf{ee},\mathrm{n},t}} \quad (4.36)$$

Sind die Daten der gesamten Sequenz bekannt, kann durch die gleichzeitige Minimierung aller Terme (engl. *batch estimation*) in obiger Gleichung ein globales Optimum des Parametersatzes $\boldsymbol{\theta}_{\mathcal{T}}$ garantiert werden.

Die rekursive Formulierung der Schätzaufgabe führt zum sog. *Kalman-Glätter*, der im linearen Fall einen optimalen Schätzer darstellt [Streit u. Luginbuhl, 1995; Logothetis u. Krishnamurthy, 1999; Al-Naffouri, 2007]. Auch für nichtlineare Systemgleichungen kann eine iterative Version des Kalman-Glätters eine optimale Schätzung zumindest annähern [Bell u. Cathey, 1993; Gauvrit u. a., 1997]. Bei der rekursiven Auswertung wird, auf der Basis des Bayes-Filters, lediglich die Marginaldichte $P(\boldsymbol{\theta}_T|\boldsymbol{\varsigma}_{1:T})$ ausgewertet, was eine effiziente rechentechnische Umsetzung erlaubt. Eine kurze Herleitung des Bayes-Filters ist in Anhang A.4.1 zu finden. Nimmt man im Fall eines normalverteilten Systemzustandes die initiale Wahrscheinlichkeitsdichte $\mathcal{N}(\boldsymbol{\theta}_0, \mathbf{P}_0)$ als bekannt an, so schätzt das Bayes-Filter den Zustand des Systems unter der Markov-Annahme für alle folgenden Zeitschritte optimal[9].

Das in dieser Arbeit entwickelte Verfahren zur schritthaltenden Schätzung der Bewegungsparameter knüpft an diesen Ansatz an und schätzt die jeweiligen Parameter auf der Basis eines Kalman-Filters in Vorwärtsrichtung durch die Zeit. Um die Nichtlinearitäten im Beobachtungsmodell zu kompensieren, wird ein erweitertes Kalman-Filter verwendet. Die Linearisierungsfehler, die aufgrund der dort durchgeführten Taylorapproximation entstehen, werden dabei kontinuierlich innerhalb der Iterationsschleife der vorgestellten Optimierungsstrategie reduziert. Mit dem entwickelten Filter konnten die Auswirkungen der Nichtlinearitäten in der Beobachtungsgleichung auf die Güte der Schätzung erheblich reduziert werden. In der Praxis hat sich gezeigt, dass der maßgebliche Teil der Verbesserung der Schätzergebnisse im Innovationsschritt bereits nach wenigen Iterationen erreicht war. Im Rückwärtszweig des klassischen Kalman-Glätters läuft das Verfahren dann durch die gesamte Sequenz der Beobachtungen zurück, womit ein globales Optimum

[9] In Anhang A.4.2 wird gezeigt, dass für das in dieser Arbeit verwendete Filter der bedingte Erwartungswert $\hat{\boldsymbol{\theta}}$ denjenigen ausgezeichneten Wert der Verteilungsdichte kennzeichnet, für den die Varianz des Schätzfehlers minimal wird. Dieser Schätzwert ist wegen der Symmetrie der Gaussverteilung identisch mit dem wahrscheinlichsten Wert des Zustands, d. h. dem MAP-Schätzwert der durch aufsuchen des Maximums der Verteilung $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = P(\boldsymbol{\theta}_T|\boldsymbol{\varsigma}_{1:T})$ gefunden wird (dieser stimmt auch mit dem der Verbundwahrscheinlichkeit $P(\boldsymbol{\theta}_{\mathcal{T}}|\boldsymbol{\varsigma}_{1:T})$ überein [Krebs, 1999].)

garantiert werden kann. Für lange, bzw. zeitlich anwachsende Messreihen wird eine solche Optimierung jedoch schnell unhandlich, da für jeden Schätzzyklus die Bestimmung der Verbundwahrscheinlichkeit $P(\boldsymbol{\theta}_{\mathcal{T}}|\boldsymbol{\zeta}_{\mathcal{T}})$ des gesamten Sequenz erforderlich ist. Diese Art der Datenauswertung ist für eine Anwendung in der mobilen Umfeldwahrnehmung praktisch nicht realisierbar und meist auch nicht notwendig. Eine Möglichkeit, die Komplexität der Schätzaufgabe zu reduzieren, ist die Betrachtung nur eines begrenzten Zeithorizonts $\mathrm{z} >= 0$ zurückliegender Beobachtungen. Je länger dieser Zeithorizont ist, desto besser nähert die resultierende Schätzung den optimalen Wert, den man bei Auswertung aller Beobachtungen ($\mathrm{z} = T$) erhält, an. Die Glättung findet somit nur bis zu den, innerhalb des Zeithorizonts $T - \mathrm{z}$ befindlichen, Beobachtungen statt. Es gilt zu bemerken, dass die beste Schätzung des Zustandes hier zum Zeitpunkt $T - \mathrm{z}$ vorliegt. Um eine möglichst zeitnahe Schätzung der gegenwärtig vorherrschenden Systemeigenschaften zu erreichen, wurde in dieser Arbeit der Zeithorizont $\mathrm{z} = 1$ gesetzt, d. h. die Auswertung umfasst zu jedem Zeitpunkt $t$ genau einen Prädiktions-Glättungs-Schritt. Da die Anzahl der Objekthypothesen innerhalb einer Iteration als konstant angenommen wird, erscheint dies für die Anwendung auf Verkehrsszenarien mit einer sich ständig ändernden Anzahl unabhängig bewegter Objekte als sinnvoll. Aufbauend auf den in Anhang A.4 hergeleiteten Gleichungen für das Kalman-Filter und das erweiterte Kalman-Filter (EKF), kann das hier entwickelte Filter in zwei Teile zerlegt werden: In Vorwärtsrichtung besteht der erste Verarbeitungsschritt aus der *Prädiktion* der Systemzustände

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{k}_{t|t-1} &= \mathbf{f}(\hat{\boldsymbol{\theta}}^{k}_{t-1}, \mathbf{u}_{t-1}, \mathbf{q}_{t-1}) = \mathbf{A}_{t-1}\hat{\boldsymbol{\theta}}^{k}_{t-1} \quad \text{und} \\ \mathbf{P}^{k}_{t|t-1} &= \mathbf{A}_{t-1}\mathbf{P}^{k}_{t-1}\mathbf{A}^{\mathrm{T}}_{t-1}\Sigma^{*}_{\mathbf{ee},t} \end{aligned} \tag{4.37}$$

in den nächsten Zeitschritt. Aufgrund der Einfachheit des verwendeten Systemmodells konstanter Beschleunigung, wird auf eine Linearisierung der Systemgleichung $\mathbf{f}$ verzichtet. Die Dynamik des Systems kann somit einfach durch die Transitionsmatrix $\mathbf{A}$ beschrieben werden. An die Prädiktion anschließend, werden die a-priori Schätzwerte mit Hilfe der aktuellen Beobachtung $\hat{\boldsymbol{\zeta}}_t$ im *Innovationschritt* korrigiert:

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{k+1}_{t} &= \hat{\boldsymbol{\theta}}^{k}_{t|t-1} + \mathbf{K}^{k}_{t}\left[\left(\hat{\boldsymbol{\zeta}}_t - \mathbf{h}(\hat{\boldsymbol{\theta}}^{k}_{t}, \mathbf{0})\right) - \mathbf{H}^{k}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}^{k}_{t|t-1} - \hat{\boldsymbol{\theta}}^{k}_{t}\right)\right] \\ \mathbf{P}^{k}_{t} &= \left[\mathrm{I} - \mathbf{K}^{k}_{t}\mathbf{H}^{k}_{\boldsymbol{\theta}}\right]\mathbf{P}^{k}_{t|t-1} \\ \mathbf{K}^{k}_{t} &= \mathbf{P}^{k}_{t|t-1}(\mathbf{H}^{k}_{\boldsymbol{\theta}})^{\mathrm{T}}\left[\mathbf{H}_{\boldsymbol{\theta}}\mathbf{P}^{k}_{t|t-1}(\mathbf{H}^{k}_{\boldsymbol{\theta}})^{\mathrm{T}} + \Sigma^{*}_{\mathbf{ee},t}\right]^{-1}. \end{aligned} \tag{4.38}$$

$\mathbf{K}_t$ drückt hierbei die Kalman-Verstärkungsmatrix aus. $\mathbf{H}_{\boldsymbol{\theta}}$ bezeichnet die Jakobi-Matrize, die sämtliche partiellen Ableitungen der nichtlinearen Beobachtungs-

funktion **h** nach den Parametern $\boldsymbol{\theta}$ enthält. In [Koch u. Yang, 1998; Younis, 1996; Ting u. a., 2007] wird eine ähnliche Beschreibung genutzt, um das jeweils vorgestellte Filter robust gegen Ausreißer zu machen.

Für die Auswertung in Rückwärtsrichtung werden die a-posteriori und a-priori Schätzwerte $\hat{\boldsymbol{\theta}}_t^{k+1}$ und $\hat{\boldsymbol{\theta}}_{t|t-1}^{k}$ sowie die entsprechenden Kovarianzmatrizen $\mathbf{P}_t^k$ und $\mathbf{P}_{t|t-1}^k$ gespeichert. Die rückwärtsgerichte Glättung des Systemzustandes erfolgt, entsprechend den Ausführungen von [Rauch u. a., 1965], nach folgendem Schema:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{t-1|T}^{k+1} &= \hat{\boldsymbol{\theta}}_{t-1}^{k} + \mathbf{G}_{t-1}^k \left[ \hat{\boldsymbol{\theta}}_{t|T}^{k+1} - \hat{\boldsymbol{\theta}}_{t|t-1}^{k} \right] \\
\mathbf{P}_{t-1|T}^k &= \mathbf{P}_{t-1}^k + \mathbf{G}_{t-1}^k \left[ \mathbf{P}_{t|T}^{k+1} - \mathbf{P}_{t|t-1}^k \right] \mathbf{G}_{t-1}^{\mathrm{T}} \\
\mathbf{G}_{t-1}^k &= \mathbf{P}_{t-1}^k \mathbf{A}_t (\mathbf{P}_{t|t-1}^k)^{-1} \, .
\end{aligned} \tag{4.39}$$

Diese Beschreibung gilt für alle $t = (T-1, T-2, \ldots, 0)$, wobei in dieser Arbeit nur eine Glättung bis $T-1$ durchgeführt wird. Die Initialisierung für die Glättung erfolgt bei $t = T$ gemäß $\hat{\boldsymbol{\theta}}_{t|T} = \hat{\boldsymbol{\theta}}_T$. $\mathbf{G}$ wird hierbei als *Glätter-Verstärkungsmatrix* bezeichnet. Diese Beschreibung ist in der Literatur unter dem Begriff *Rauch-Tung-Striebel*-(RTS-)Glätter bzw. *Alpha-Gamma*-Algorithmus bekannt. Eine ähnliche Formulierung wird von [Jazwinski, 1970] als *iterated filter smoother* und von [Wishner u. a., 1969] als *single stage interation filter* bezeichnet. Die prinzipielle Funktionsweise des Verfahrens ist in Tabelle 4.1 nochmals zusammenfassend dargestellt. Bei der Implementierung des Verfahrens zeigt sich, dass der Iterationsprozess bereits bei sehr kleinen Werten konvergiert. In dieser Arbeit war der größte Teil der Verbesserungen bereits nach $K = 3-5$ Iterationsschritten erreicht, womit der Rechenaufwand je Iterationszyklus vertretbar ist. Um den Rechenaufwand zu reduzieren, wurden auch Untersuchungen durchgeführt, die Kalman-Verstärkungsmatrix und die Kovarianzmatrix innerhalb des Iterationszyklus konstant zu halten.

Durch die oben vorgestellte Optimierungsstrategie werden bei jedem Iterationsschritt $k$ die Modellparameter approximativ bestimmt. Darauf aufbauend, wird die bedingte Log-Likelihood der vollständigen Daten ausgewertet. Dieser Vorgang wird solange wiederholt, bis die Änderungen $|\mathrm{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^k) - \mathrm{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{k+1})|$ zweier aufeinander folgender Schritte unter eine vordefinierte Schwelle fällt, oder eine maximale Anzahl an Iterationsschritten erreicht wird.

(I) **Initialisierung** zum aktuellen Zeitpunkt $T = 2$

$$\boldsymbol{\theta}_{1|0} \sim \mathcal{N}(0, \mathbf{P}_0), \mathbf{P}_{1|0} = \mathbf{P}_0,$$

(II) **Iterationsschleife** $k = 1, \ldots, K$

**Filterschritt** von $t = T - \mathsf{z}, \ldots, T$

Prädiktion

$$\hat{\boldsymbol{\theta}}^{k}_{t|t-1} = \mathbf{A}_{t-1} \hat{\boldsymbol{\theta}}^{k}_{t-1}$$

$$\mathbf{P}_{t|t-1} = \mathbf{A}_{t-1} \mathbf{P}_{t-1} \mathbf{A}^{\mathrm{T}}_{t-1} \Sigma^{*}_{\mathbf{ee},t}$$

Innovation

$$\hat{\boldsymbol{\theta}}^{k+1}_{t} = \hat{\boldsymbol{\theta}}^{k}_{t|t-1} + \mathbf{K}^{k}_{t} \left[ \left( \hat{\boldsymbol{\varsigma}}_t - \mathbf{h}(\hat{\boldsymbol{\theta}}^{k}_{t}, \mathbf{0}) \right) - \mathbf{H}^{k}_{\boldsymbol{\theta}} \left( \hat{\boldsymbol{\theta}}^{k}_{t|t-1} - \hat{\boldsymbol{\theta}}^{k}_{t} \right) \right]$$

$$\mathbf{P}^{k}_{t} = \left[ \mathrm{I} - \mathbf{K}^{k}_{t} \mathbf{H}^{k}_{\boldsymbol{\theta}} \right] \mathbf{P}^{k}_{t|t-1}$$

$$\mathbf{K}^{k}_{t} = \mathbf{P}^{k}_{t|t-1} (\mathbf{H}^{k}_{\boldsymbol{\theta}})^{\mathrm{T}} \left[ \mathbf{H}_{\boldsymbol{\theta}} \mathbf{P}^{k}_{t|t-1} (\mathbf{H}^{k}_{\boldsymbol{\theta}})^{\mathrm{T}} + \Sigma^{*}_{\mathbf{ee},t} \right]^{-1}$$

Ausgabe: $\hat{\boldsymbol{\theta}}^{k+1}_{T} = \hat{\boldsymbol{\theta}}_{T|T}$

**Glättungsschritt** von $t = T, \ldots, T - \mathsf{z}$

$$\hat{\boldsymbol{\theta}}^{k+1}_{t-1|T} = \hat{\boldsymbol{\theta}}^{k}_{t-1} + \mathbf{G}^{k}_{t-1} \left[ \hat{\boldsymbol{\theta}}^{k+1}_{t|T} - \hat{\boldsymbol{\theta}}^{k}_{t|t-1} \right]$$

$$\mathbf{P}^{k}_{t-1|T} = \mathbf{P}^{k}_{t-1} + \mathbf{G}^{k}_{t-1} \left[ \mathbf{P}^{k+1}_{t|T} - \mathbf{P}^{k}_{t|t-1} \right] \mathbf{G}^{\mathrm{T}}_{t-1}$$

$$\mathbf{G}^{k}_{t-1} = \mathbf{P}^{k}_{t-1} \mathbf{A}_t \left( \mathbf{P}^{k}_{t|t-1} \right)^{-1}$$

Ausgabe: $\hat{\boldsymbol{\theta}}^{k+1}_{T:(T-\mathsf{z})} = \left( \hat{\boldsymbol{\theta}}^{k+1}_{T}, \ldots, \hat{\boldsymbol{\theta}}^{k+1}_{T-\mathsf{z}} \right)$

(III) **Nächster Zeitschritt**

setze $T := T + 1$, gehe zurück zu (II)

**Tabelle 4.1**: Funktionsweise des Kalman Glätters mit Zeithorizont $\mathsf{z} = 1$ und zeitlich anwachsender Messreihe.

## 4.4 Hypothesenverwaltung

Eine wichtige Größe bei der hier vorgestellten Szenensegmentierung stellt die Anzahl $J$ der unabhängig bewegten Objekte in der Szene dar. Hierbei müssen die initialisierten Objektinstanzen auf ihre Gültigkeit hin validiert und potentielle neue Objekthypothesen dem Prozess hinzugefügt werden. Hierzu wird die aktuelle Segmentierung genutzt, wie Abbildung 4.4 verdeutlicht. Bildpunkte, die einen hohen Segmentierungsfehler[10] $C_{\mathrm{E}}$ erzeugen, werden hier dem Null-Label zugewiesen.

[10] Die Definition des Gütemaßes findet sich in (5.1) im Ergebnisteil dieser Arbeit.

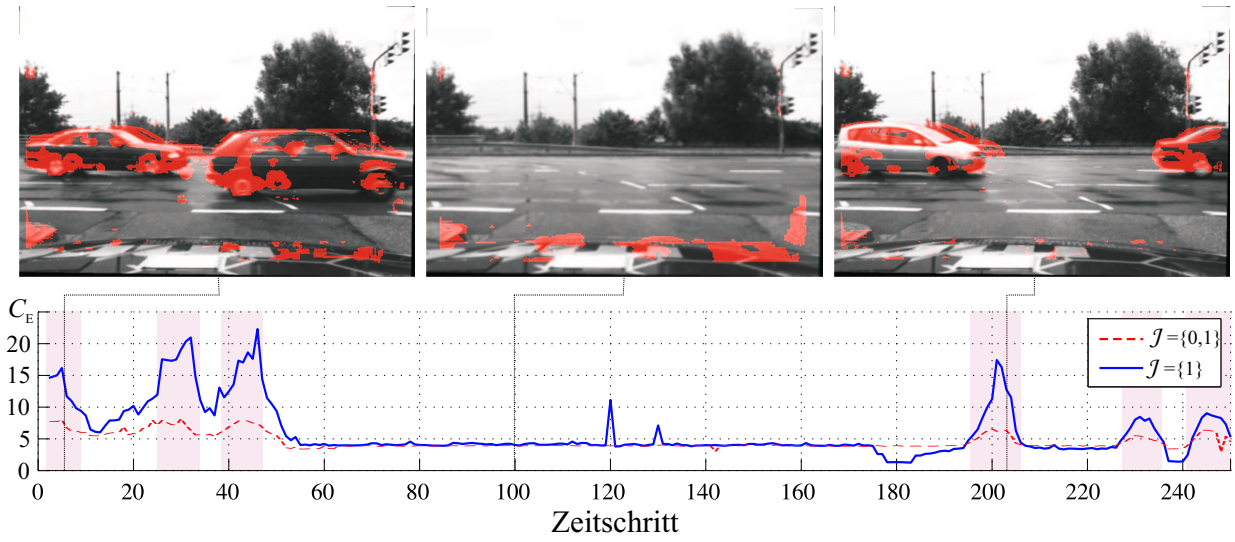



**Abbildung 4.4**: Oben: Momentaufnahmen einer Hintergrundsegmentierung mit mehreren bewegten Objekten. Unten: Die blaue Kurve zeigt den Restfehler $C_E$ der Segmentierung, wobei das Null-Label nicht in der Labeldefinition berücksichtigt ist, d. h. $\mathcal{J} = \{1\}$. Punkte, die auf unabhängig bewegten Objekten liegen, erzeugen somit einen hohen Fehlereintrag in der Restfehlerkarte, was in einem Anstieg des Gesamtrestfehlers erkennbar ist. Die entsprechenden Bereiche sind farblich markiert. Die rote Kurve zeigt den Restfehler für den Standardfall, in dem das Null-Label bei der Segmentierung berücksichtigt wird, d. h. $\mathcal{J} = \{0, 1\}$. In diesem Fall wird Bildpunkten, die nicht eindeutig der Eigenbewegung zugewiesen werden können, das Null-Label aufgeprägt. In den Bildern im oberen Teil der Abbildung sind diese Bereiche rot markiert.

Merkmalspunkte, die in so klassifizierten Bildbereichen liegen, werden auf ihre Bewegungsähnlichkeit zueinander überprüft und im zutreffenden Fall ähnlicher Bewegung und Lage im Raum zur Initialisierung einer neuen Hypothese genutzt.

### 4.4.1 Erzeugung von Hypothesen

Durch die Auswertung der Merkmalspunkte bzgl. einer gemeinsamen charakteristischen Bewegung und räumlicher Konsistenz, kann effizient nach Bewegungsmustern gesucht werden, die auf noch nicht detektierte, unabhängig bewegte Objekte in der Szene hinweisen. Hierfür wurden zwei Verfahren entwickelt, die im Folgenden näher vorgestellt werden.

### Homographiebasiertes Verfahren

Zur Identifikation von Gruppierungen ähnlich starr bewegter Punkte wird ein iterativer RANSAC-Ansatz (engl. *RANSAC=Random Sample Consensus*) [Fischler u. Bolles, 1981] verwendet [Bachmann u. Kuehne, 2009]. Hierbei werden Punktmengen zu Gruppen zusammengefasst, die über eine bestimmte Zeitdauer von $m$ Bildern durch das Modell gleicher konstanter Bewegung beschrieben werden können. Jede der Gruppierungen umfasst eine Menge zufällig gezogener Punkte $\mathcal{X}_s$, mit denen die 2D Homografie zu allen $m-1$ korrespondierenden Punkten berechnet wird. Der Abstand von jeder möglichen Korrespondenz in der Menge aller Punkte wird durch einen symmetrischen Transformationsfehler beschrieben [Hartley u. Zisserman, 2004]. Um lokale Zusammengehörigkeit und Bewegungsähnlichkeit der Punktmenge zu gewährleisten, wird die Distanzfunktion mit der Distanz aus räumlichem Abstand und Differenz der Bewegungsvektoren gewichtet. Es wird angenommen, dass Punkte zur gleichen Gruppe gehören, deren Distanz zu der geschätzten Homografie unterhalb eines vorher festgelegten Schwellwerts liegt. Punkte, die unterhalb dieses Abstands liegen, werden zu einer potentiellen neuen Objekthypothese zusammengefasst und aus der Liste der zu untersuchenden Punkte gelöscht. Dieser Vorgang wird wiederholt, bis die Anzahl verbleibender Punkte unterhalb eine bestimmten Grenze fällt. Aufgrund der Auswertung mehrerer Bilder gleichzeitig, steht dieser Ansatz jedoch hinter dem hierarchischen Ballungsansatz (engl. *hierarchical clustering*) zurück, welcher bei wesentlich geringerem Rechenaufwand vergleichbare Ergebnisse liefert.

### Hierarchisches Ballungsverfahren

Als Merkmal wird hier die Position der 3D Punkte in der Szene verwendet. Das Kriterium zum Ballen der Punkte – im Weiteren wird hierfür auch der etwas kompaktere Begriff des „clusterns" verwendet – ist der räumliche Abstand zwischen den einzelnen Punkten. Punkte, deren Abstand einen Schwellwert unterschreiten, werden dabei zu einem Cluster zusammengefasst [Kitt, 2008].

Beim hierarchischen Clustern wird eine Pyramidenstruktur verwendet. Jede Ebene dieser Struktur beinhaltet eine gewisse Anzahl erzeugter Cluster, die die zusammengehörigen Merkmalsvektoren beinhalten [Elias, 2003]. Die Basis der Pyramide besteht aus einer Menge von Punkten, welche als fremdbewegte Raumpunkte identifiziert wurden. Mit jeder Ebene werden die Raumpunkte einer stetig kleiner werdenden Zahl von Clustern zugewiesen. An der Spitze der Pyramide verbleiben jeweils die Raumpunkte in einem einzelnen Cluster, die aufgrund ihrer Distanz zueinander als räumlich zusammengehörend betrachtet werden können. Die Erzeugung neuer Ebenen der Pyramide sowie die Verbindungen der jeweiligen Cluster

innerhalb einer Ebene sind durch mehrere Regeln definiert, die im Folgenden näher erläutert werden. Für die Erzeugung neuer Cluster wird eine binäre Vergleichsvariable $q$ definiert, die festlegt, ob zwei Cluster innerhalb einer Ebene einen bestimmten Abstand unterschreiten oder nicht. Wird der Abstand unterschritten, so wird diese Variable aktiviert, d. h. auf den Wert $q = 1$ gesetzt. Wird der Abstand nicht unterschritten, ist die Vergleichsvariable inaktiv, was einem Wert von $q = 0$ entspricht. Ist letzteres der Fall, werden beide Cluster unverändert in die nächste Ebene übernommen. Ist die Variable $q = 1$, werden die beiden Cluster zu einem neuen Cluster zusammengefasst. Nach [Elias, 2003] können folgende beiden Regeln definiert werden:

- ◇ Zwei Cluster überleben genau dann den Übergang in eine höhere Ebene, wenn die binäre Vergleichsvariable zu 0 gesetzt wurde.
- ◇ Für jedes Cluster, das beim Übergang in die nächsthöhere Ebene eliminiert wird, existiert mindestens ein benachbartes Cluster, das diesen Übergang überlebt.

Es wird angenommen, dass die Ebene $e$ der Pyramide durch eine Menge $\mathcal{M}_e = \{\mathcal{O}_{e,1}, \mathcal{O}_{e,2}, \ldots, \mathcal{O}_{e,c}\}$ an Clustern $\mathcal{O}_{e,c}$ gegeben ist. $c$ gibt hier die Anzahl an Clustern in der Ebene $e$ an. Jedes dieser Cluster $\mathcal{O}_{e,u}$ besteht aus einer Menge von Punkten $\{\mathbf{X}\}_{w=1}^{M_{e,u}}$, wobei $M_{e,u}$ die Anzahl der Punkte in einem Cluster angibt.

Zum Erzeugen neuer Ebenen wird für jedes mögliche Clusterpaar $\mathcal{O}_{e,u}$ und $\mathcal{O}_{e,v}$ innerhalb der Ebene $e$ die binäre Vergleichsvariable $q$ zu Null gesetzt. Im Anschluss daran wird der paarweise Abstand zweier Cluster bestimmt. Hierzu können folgende Distanzmaße genutzt werden:

(i) Der *geringste Abstand zwischen Punkten* innerhalb zweier Cluster

$$d(u,v) = \min \left\{ \operatorname{dist}\left( \{\mathbf{X}\}_{w=1}^{M_{e,u}}, \{\mathbf{X}\}_{s=1}^{M_{e,v}} \right) \right\} . \tag{4.40}$$

(ii) Der *mittlere Abstand aller möglichen Punktepaare* der beiden Cluster

$$d(u,v) = \frac{1}{M_{e,u} M_{e,v}} \sum_{\mathbf{X}_w \in \mathcal{O}_{e,u}} \sum_{\mathbf{X}_s \in \mathcal{O}_{e,v}} \operatorname{dist}(\mathbf{X}_w, \mathbf{X}_s) . \tag{4.41}$$

(iii) Der *euklidische Abstand der Zentren* beider Cluster

$$\begin{aligned} d(u,v) &= \left\| \overline{\mathbf{X}}_{e,u} - \overline{\mathbf{X}}_{e,v} \right\| , \text{mit} \\ \overline{\mathbf{X}}_{e,u} &= \frac{1}{M_{e,u}} \sum_{\mathbf{X}_w \in \mathcal{O}_{e,u}} \mathbf{X}_w \quad \text{und} \quad \overline{\mathbf{X}}_{e,v} = \frac{1}{M_{e,v}} \sum_{\mathbf{X}_s \in \mathcal{O}_{e,v}} \mathbf{X}_s . \end{aligned} \tag{4.42}$$

| q | $\mathcal{O}_{e+1,u}$ existiert: J | $\mathcal{O}_{e+1,u}$ existiert: N | $\mathcal{O}_{e+1,v}$ existiert |
|---|---|---|---|
| 1 | $\mathcal{O}_{e,u} \rightarrowtail \mathcal{O}_{e+1,v}$, $\mathcal{O}_{e+1,u} \rightarrowtail$ del<br>$\mathcal{O}_{e,v} \rightarrowtail \mathcal{O}_{e+1,u}$, $\mathcal{O}_{e+1,v} \rightarrowtail$ del | $\mathcal{O}_{e,u} \rightarrowtail \mathcal{O}_{e+1,v}$ | J |
| 1 | $\mathcal{O}_{e,v} \rightarrowtail \mathcal{O}_{e+1,u}$ | $\mathcal{O}_{e,u} \rightarrowtail \mathcal{O}^{\text{neu}}_{e+1,w}$<br>$\mathcal{O}_{e,v} \rightarrowtail \mathcal{O}^{\text{neu}}_{e+1,w}$ | N |
| 0 | - | $\mathcal{O}_{e,u} \rightarrowtail \mathcal{O}^{\text{neu}}_{e+1,u}$ | J |
| 0 | $\mathcal{O}_{e,v} \rightarrowtail \mathcal{O}^{\text{neu}}_{e+1,v}$ | $\mathcal{O}_{e,v} \rightarrowtail \mathcal{O}^{\text{neu}}_{e+1,v}$<br>$\mathcal{O}_{e,u} \rightarrowtail \mathcal{O}^{\text{neu}}_{e+1,u}$ | N |

**Abbildung 4.5**: Fallunterscheidungen des hierarchischen Ballungsverfahrens. „del" steht für die Vernichtung des entsprechenden Clusters, „neu" für eine Neuinitialisierung. „J" für Ja und „N" für Nein geben an, ob ein übergeordneter Elternknoten existiert.

Mit Hilfe der so bestimmten Abstände wird der Wert der Vergleichsvariable $q$ dann folgendermaßen bestimmt:

$$q = \begin{cases} 0 & \text{für } d(u,v) \geq T_{\mathcal{O}} \\ 1 & \text{für } d(u,v) < T_{\mathcal{O}}\,. \end{cases} \tag{4.43}$$

Um anhand der Vergleichsvariablen $q$ neue Cluster zu erzeugen, müssen verschiedene Fälle unterschieden werden, die in Abbildung 4.5 grafisch aufbereitet sind. Dieser Algorithmus wird iterativ so lange durchlaufen, bis an der Spitze der Pyramide nur noch Cluster vorhanden sind, deren paarweiser Abstand untereinander den Schwellwert $T_{\mathcal{O}}$ überschreitet, sodass keine neuen Cluster mehr gebildet werden können. Mit den so gefundenen Merkmalsgruppierungen werden dann die initialen Parameter der potentiellen Objekthypothesen geschätzt.

In der Praxis hat sich gezeigt, dass der hierarchische Clusteransatz in Kombination mit einem Plausibilitätstest gute Ergebnisse erzeugt. Die Plausibilitätsprüfung umfasst hier u. a. die Untersuchung der Höhe des Schwerpunkts relativ zur geschätzten Fahrbahnebene sowie Ausdehnung und Kompaktheit der Merkmalspunkte einer potentiellen Hypothese.

### 4.4.2 Vernichtung von Hypothesen

Die Vernichtung einer Hypothese ist bereits durch die Segmentierung selbst gegeben. Werden einem Hypothesenlabel keine Punkte zugewiesen, d. h. $\sum_{n\in\mathcal{R}} \hat{l}_n^j = 0$, wird die Hypothese gelöscht.

## 4.5 Zusammenfassung

Zur Lösung des gekoppelten Optimierungsproblems der dichten Szenensegmentierung wird die Schätzung der Objektparameter als Problem unvollständiger Daten beschrieben und in alternierender Reihenfolge, ähnlich dem bekannten EM-Verfahren, gelöst. Dabei werden die Beobachtungen nicht entsprechend einer binären Entscheidungsregel nur einer bestimmten Objekthypothese zugewiesen, sondern probabilistisch gewichtet auf die Menge aller Objekthypothesen verteilt. Die Glattheitsanforderung einzelner Objekte wird durch die Modellierung des Assoziationsprozesses als Markov-Zufallsfeld berücksichtigt. Zur Schätzung der Bewegungsparameter wird ein iteratives erweitertes Kalman-Filter eingeführt, welches die Wahrscheinlichkeit der einzelnen Beobachtungen auf die Menge der vorhandenen Objekthypothesen abbildet und entsprechend bei der Schätzung berücksichtigt. Die Erzeugung und Vernichtung der im Schätzprozess enthaltenen Objekthypothesen erfolgt merkmalsbasiert und ist gekoppelt mit dem aktuellen Segmentierergebnis.

KAPITEL 5

# Experimentelle Auswertung

In diesem Kapitel werden die Ergebnisse der Szenensegmentierung am Beispiel typischer Verkehrsszenarien gezeigt. Die Leistungsfähigkeit des Verfahrens wurde dabei anhand von realen und synthetisch generierten Stereobildsequenzen untersucht und ausgewertet. Abbildung 5.1 zeigt beispielhaft eine Auswahl der in dieser Arbeit verwendeten Bildfolgen. Neben der Auswertung der Teilaufgaben Szenenrekonstruktion, Bewegungschätzung und Bildsegmentierung unabhängig voneinander

**Abbildung 5.1**: Momentaufnahmen einiger in dieser Arbeit verwendeten Stereosequenzen. Die realen Bilddaten wurden mit den Versuchsträgern des Instituts für Mess- und Regelungstechnik (MRT) aufgenommen. Hierfür liegen größtenteils die Informationen der Eigenfahrzeugdynamik in Form von IMU/GPS-Daten vor. Diese wurden u. a. bei der Auswertung der Güte der Bewegungsschätzung verwendet.

ander, wird auch der Einfluss der einzelnen Schätzgrößen auf das Gesamtergebnis untersucht und bewertet. Zusätzlich zur Bewertung der Szenensegmentierung anhand verschiedener Gütemaße werden repräsentative Segmentierergebnisse visualisiert und dem Leser für eine qualitative Bewertung zugänglich gemacht.

## 5.1 Rekonstruktion der 3D Szene

Die Güte der Szenenrekonstruktion wurde anhand von synthetisch generierten Bildsequenzen[1] ausgewertet, für die entsprechende Referenzdaten in Form dichter Disparitätskarten vorlagen. Zur Bewertung der Verfahren wurde das in (3.27) definierte Gütemaß verwendet. Neben dem Rekonstruktionsfehler wurde auch der Segmentierungsfehler der resultierenden Szenenzerlegung bei der Bewertung des Verfahrens mit berücksichtigt. Die für eine Segmentierung benötigte Bewegungsinformation wurde hierbei nicht geschätzt, sondern als fehlerfrei bekannt vorausgesetzt. Um das Gütemaß auch zur Bewertung der Segmentierergebnisse mit realen Bilddaten nutzen zu können, wurde das Gütemaß durch den mittleren Restfehler

$$C_{\mathrm{E}} = \frac{1}{N} \sum_{n \in \mathcal{R}} |\varepsilon_n| \tag{5.1}$$

definiert. Da die Qualität des Segmentierergebnisses zusätzlich von der Abdeckung der Szene durch die entsprechenden Tiefenwerte abhängt, muss das Verfahren auch hinsichtlich dieses Kriteriums bewertet werden. Hierfür wurde das Verhältnis $\Gamma = N_{\mathrm{EST}}/N_{\mathrm{GT}}$ der insgesamt rekonstruierten Szenenpunkte $N_{\mathrm{EST}}$ zur tatsächlichen Anzahl von Szenenpunkte $N_{\mathrm{GT}}$ gewählt.

Die global optimierenden Verfahren wurden dabei mit einem lokalen Blockzuordnungsverfahren verglichen. Zum Vergleich der beiden Ansätze, sind in Abbildung 5.2 die Rekonstruktionsergebnisse der beiden Verfahren gegenübergestellt.

In Abbildung 5.3 sind die Ergebnisse, basierend auf den oben eingeführten Bewertungsmaßen, gezeigt. Wie zu erwarten, ist der lokale Ansatz den global optimierenden Verfahren bzgl. Laufzeit deutlich überlegen. Vergleicht man jedoch die Abdeckung der Szene durch die Tiefenkarte, zeigen sich die Vorteile eines global

[1] Ein Teil der synthetisch Bilddaten, einschließlich der Referenzdaten, ist unter `http://www.mi.auckland.ac.nz` im Rahmen des „.enpeda..”-Projekts öffentlich zugänglich.

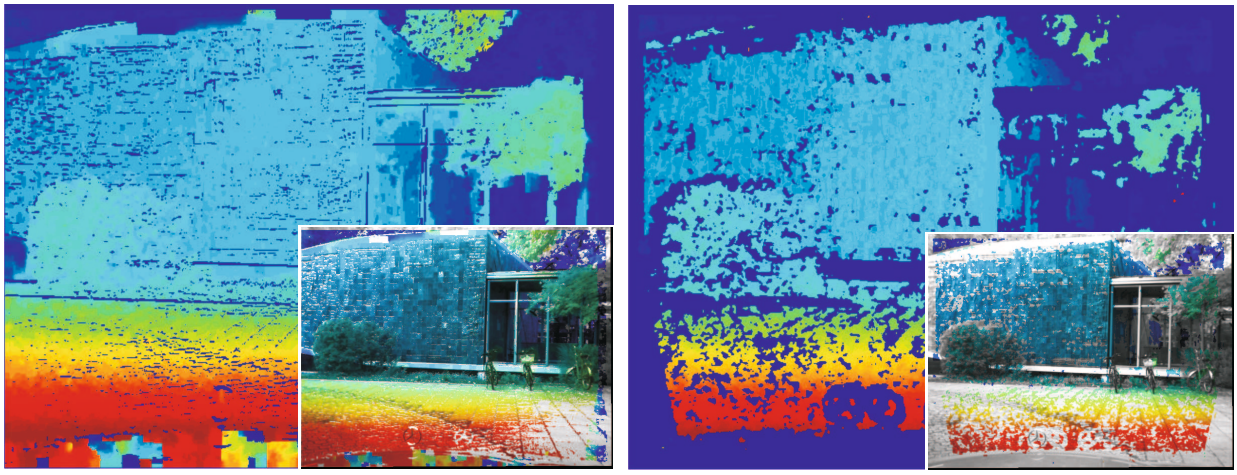


**Abbildung 5.2**: Rekonstruktionsergebnis in Form einer Disparitätskarte. Rote Farben symbolisieren Nähe; mit zunehmender Entfernung gehen die Farben ins Blaue über. In dunkelblauen Bereichen konnte keine Tiefe bestimmt werden. Links: Stereorekonstruktion mit dem global definierten Modell aus Kapitel 3. Rechts: Das Ergebnis des lokalen Blockzuordnungsverfahrens aus [Dang, 2007].

definierten Gütemaßes, welches sich über den gesamten Konfigurationsraum der Zufallsvariable erstreckt. In Bereichen, in denen die Beobachtungen selbst keine eindeutige Zuweisung erlauben, kann durch die Glattheitsannahme eine plausible Szenenoberfläche rekonstruiert werden, was zu einer fast vollständigen Abdeckung der Szene führt. Auch bzgl. Rekonstruktions- und Segmentierungsgüte erzielt die globale Szenenrekonstruktion bessere Ergebnisse im Vergleich mit dem lokalen Verfahren. Der Einfluss der zeitlichen Bindung der Tiefenwerte muss bzgl. der Rekonstruktionsgüte und Abdeckung als gering eingestuft werden. Jedoch hat die Berücksichtigung der bewegungsprädizierten Tiefenwerte eine merkliche Auswirkung auf die Laufzeit des Verfahrens, wie im unteren Teil von Abbildung 5.3 gezeigt wird. Durch die Reduktion des Suchraums auf eine plausible Untermenge an möglichen Tiefenwerten, kann so der Rechenaufwand des Verfahrens um bis zu 20% bzgl. der ursprünglichen Rechenzeit reduziert werden.

Abbildung 5.4 zeigt den Einfluss des Rekonstruktionsfehlers auf die Szenensegmentierung. Hierbei wurden die Bilddaten der linken Kamera künstlich mit additivem Rauschen überlagert. In der Abbildung ist der Rekonstruktionsfehler für ansteigende Werte des Rauschterms aufgetragen. Ab einem Rekonstruktionsfehler von $C_\Delta \approx 4,0$ ist ein deutlicher Anstieg des Segmentierungsfehlers erkennbar. Die anscheinend stagnierenden Werte von $C_\mathrm{E}$ trotz des zunehmenden Fehlers bei der Rekonstruktion kann dadurch erklärt werden, dass eine immer größere Anzahl an Bildpunkten dem Null-Label zugewiesen werden. Da der Fehler dieser

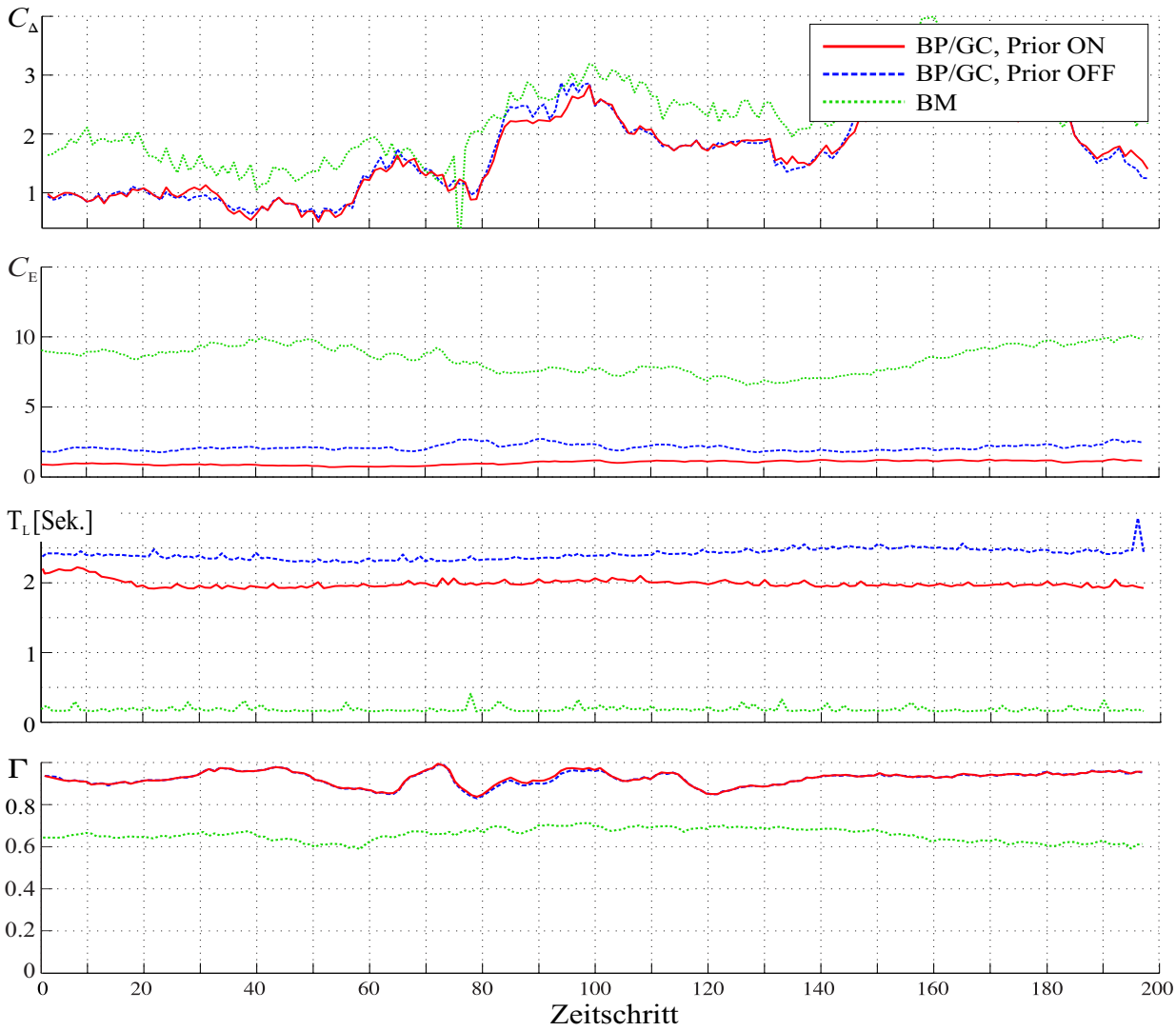



**Abbildung 5.3**: Von oben nach unten: Vergleich des mittleren Rekonstrukionsfehlers $C_\Delta$ für den Fall einer Standard-Stereorekonstruktion („Prior OFF") und unter Berücksichtigung des Prädiktionsterms aus (3.24) („Prior ON"). Zum Vergleich sind die Ergebnisse des lokalen Verfahrens (BM) mit aufgeführt. Darunter ist der Einfluss des Rekonstruktionsfehlers auf die Segmentierung in Form des mittleren Restfehlers $C_E$ gezeigt. Weiterhin sind die entsprechenden Laufzeiten $T_L$ der Rekonstruktionsverfahren und die Abdeckung $\Gamma$ der Szene durch die Tiefenkarte abgebildet.

Punkte nicht in das Gütemaß mit eingeht, kann es sogar dazu kommen, dass $C_E$ trotz großer Fehler bei der Rekonstruktion beginnt abzunehmen. Dieser Effekt tritt jedoch erst bei Werten von $C_\Delta$ auf, die in der Praxis unrealistisch sind, bzw. bei denen die Segmentierung zu augenscheinlich völlig unbrauchbaren Ergebnissen führt.

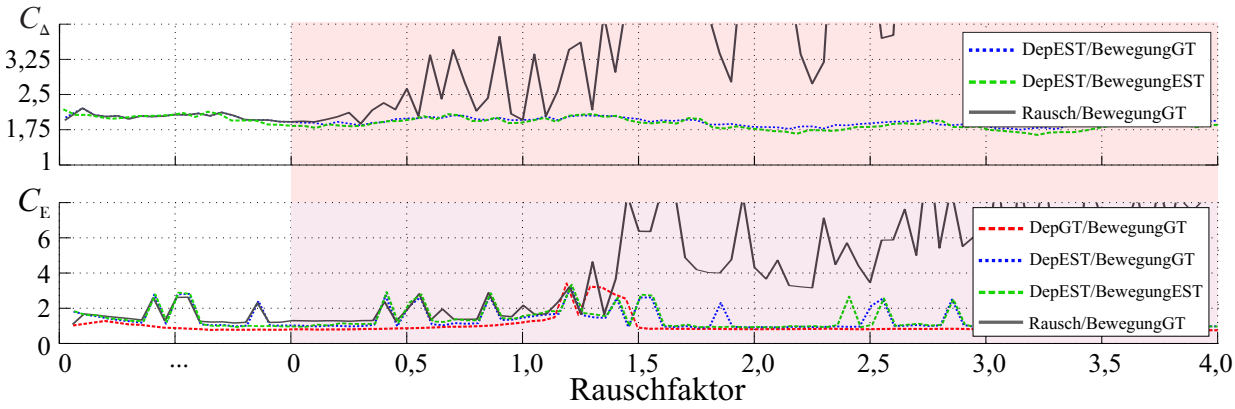



**Abbildung 5.4**: Der Einfluss des Rekonstruktionsfehlers $C_\Delta$ auf die Szenensegmentierung. Oben: Rekonstruktionsfehler $C_\Delta$ für zunehmend verrauschte Bilddaten. Der Bereich, ab dem die linke Kamera mit additivem Rauschen überlagert wird, ist rot markiert. Unten: Resultierender Segmentierungsfehler $C_E$. Hierbei wird die Bewegung als bekannt und fehlerfrei angenommen.

## 5.2 Schätzung der 3D Bewegung

Aufgrund der begrenzten Verfügbarkeit von Referenzdaten für die Bewegungsprofile fremdbewegter Objekte in den vorhandenden Bildsequenzen, beschränkt sich die quantitative Bewertung der Bewegungsschätzung größtenteils auf die Eigenbewegung des Fahrzeugs. Bei den realen Bildsequenzen ist der Schätzfehler der Eigenbewegung definiert als die relative Abweichung der jeweiligen Bewegungskomponente zu den IMU/GPS-Messungen der Fahrzeugbewegung. Im Folgenden sind die Unterbrechungen der Kennlinien in den einzelnen Abbildungen durch gelegentlich auftretende Ausfälle der Sensordatenaufzeichnung zu erklären.

Für eine reale Bildsequenz zeigt Abbildung 5.5 die Entwicklung des Schätzfehlers der einzelnen Bewegungsparameter über der Zeit aufgetragen. Zu jedem Zeitpunkt iteriert das Verfahren über eine konstante Anzahl von $K = 20$ Iterationsschritten. Zusätzlich zu den Schätzfehlern ist der Restfehler der resultierenden Segmentierung im unteren Teil der Abbildung mit angegeben. Es ist deutlich sichtbar, dass die Geschwindigkeitsschätzung innerhalb eines Iterationszyklus jeweils gegen einen bestimmten Wert konvergiert. Der Grund für die bei einzelnen Iterationsdurchläufen auftretende Zunahme des Fehlers kann hierbei nicht näher erklärt werden. Über die Zeit konvergiert die Geschwindigkeitsschätzung jedoch zu kleinen Werten und führt zu einer kontinuierlichen Abnahme des Restfehlers der Segmentierung. Die Untersuchungen haben ergeben, dass eine maximale Iterationszahl $K = 5$ im Allgemeinen zu befriedigenden Ergebnissen bei der Szenensegmentierung führt. Über längere Bildsequenzen hinweg, bricht das Verfahren auf-

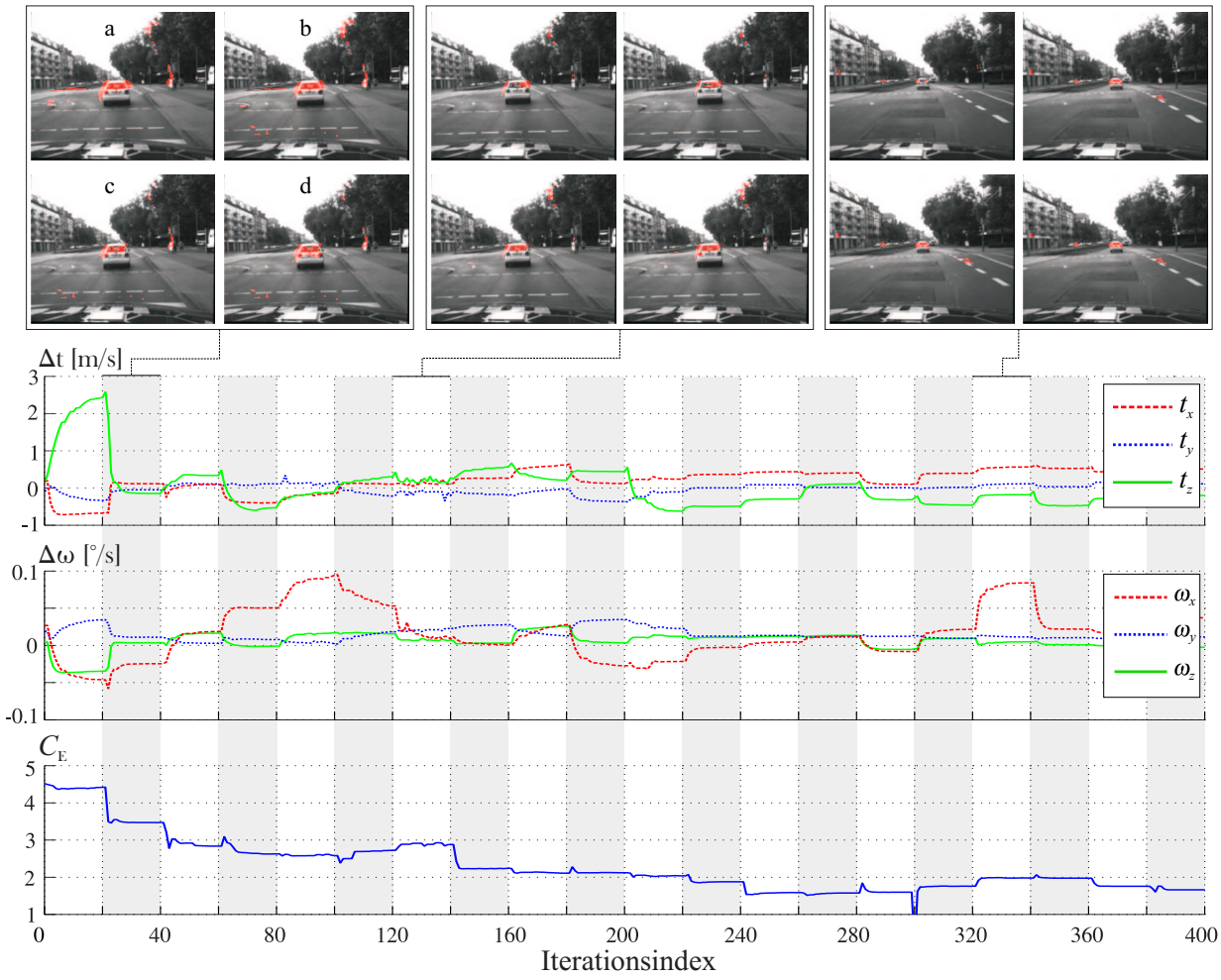



**Abbildung 5.5**: Oben: Momentaufnahmen der Segmentierung. Die einzelnen Bilder sind in alphabetisch aufsteigender Reihenfolge, wie bei der ersten Bildfolge oben links gezeigt, geordnet. Als fremdbewegt klassifizierte Szenenpunkte sind rot eingefärbt. Unten: Zeitliche Entwicklung des Fehlers der Bewegungsschätzung relativ zu den zeitsynchronen IMU/GPS-Messungen des Fahrzeugs.

grund der sich minimal ändernden Schätzgrößen meist nach $k \approx 3$ Iterationen ab. Abbildung 5.6 zeigt das Ergebnis der Bewegungsschätzung für eine reale Bildsequenz und den resultierenden Restfehler über der Zeit.

Der Einfluss des Schätzfehlers der Bewegungskomponenten auf den Segmentierprozess wird in Abbildung 5.7(a)- 5.7(b) illustriert. Hierbei wurden neben dem Fehlereinfluss jeder einzelnen Komponente auf den resultierenden Restfehler auch der realistische Fall mehrerer fehlerbehafteter Bewegungskomponenten untersucht.

Abbildung 5.8 zeigt den Segmentierungsfehler in Abhängigkeit des Schätzfehlers der Bewegung. Hierfür wurde das wahre Bewegungsprofil mit einem additiver Rauschterm überlagert. Es zeigt sich, dass der Restfehler ab einem Wert von ca.

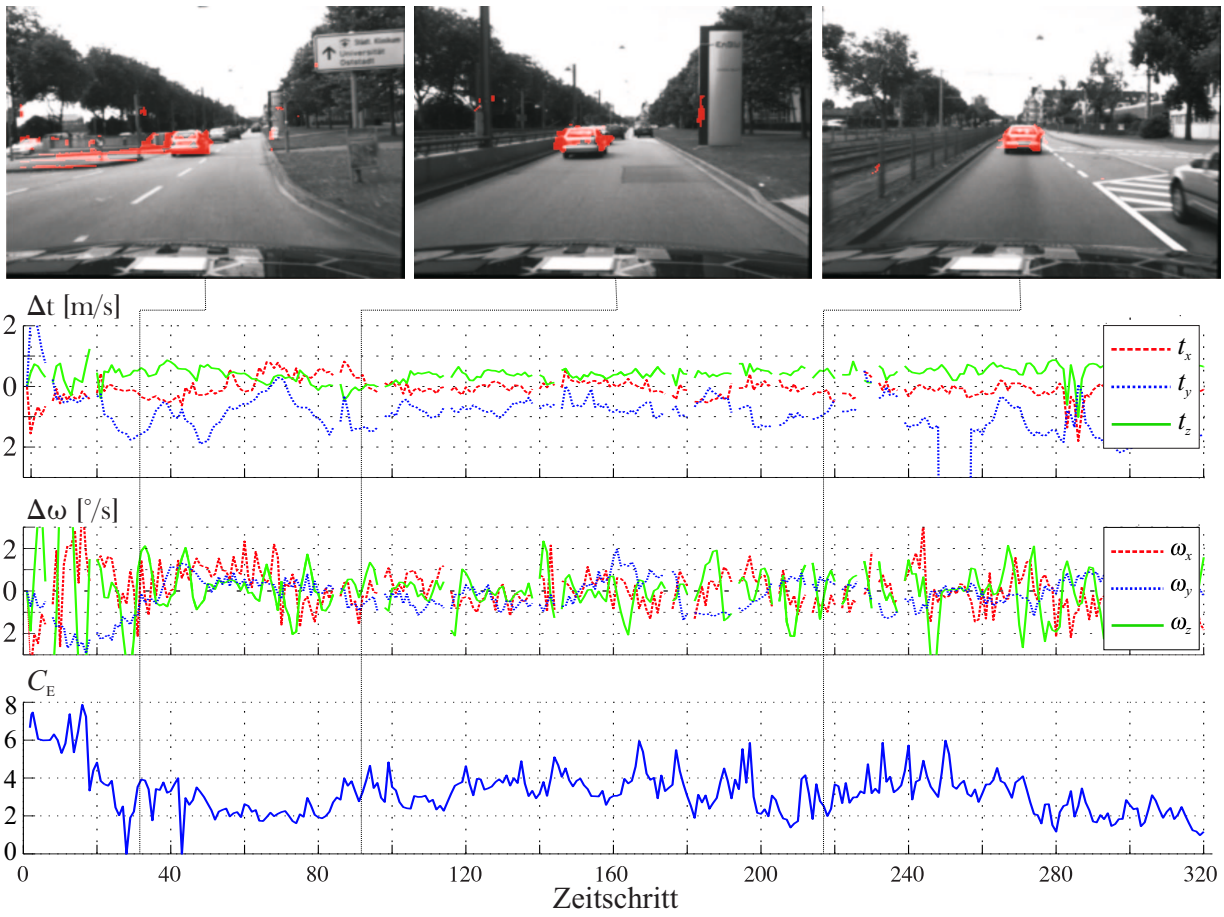
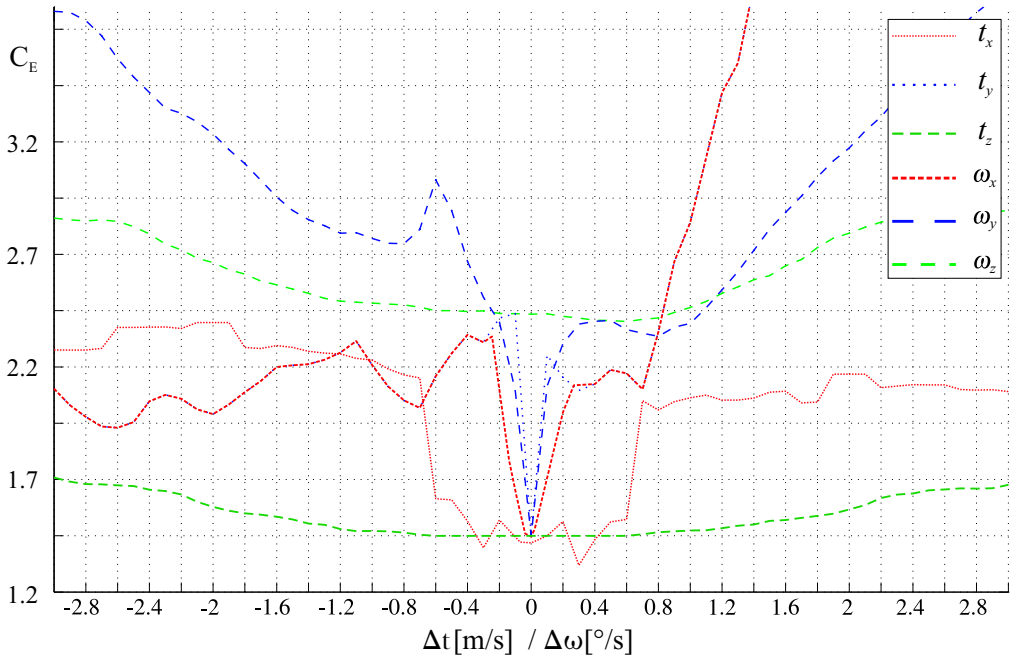



**Abbildung 5.6**: Oben: Momentaufnahmen der Szenensegmentierung. Unten: Fehler der Bewegungsschätzung relativ zur gemessenen Eigenbewegung der Inertialsensorik. Die Unterbrechungen der Kurven entstehen durch Ausfälle bei der Datenaufzeichnung. Das Diagramm im unteren Teil zeigt den resultierenden Restfehler $C_E$.

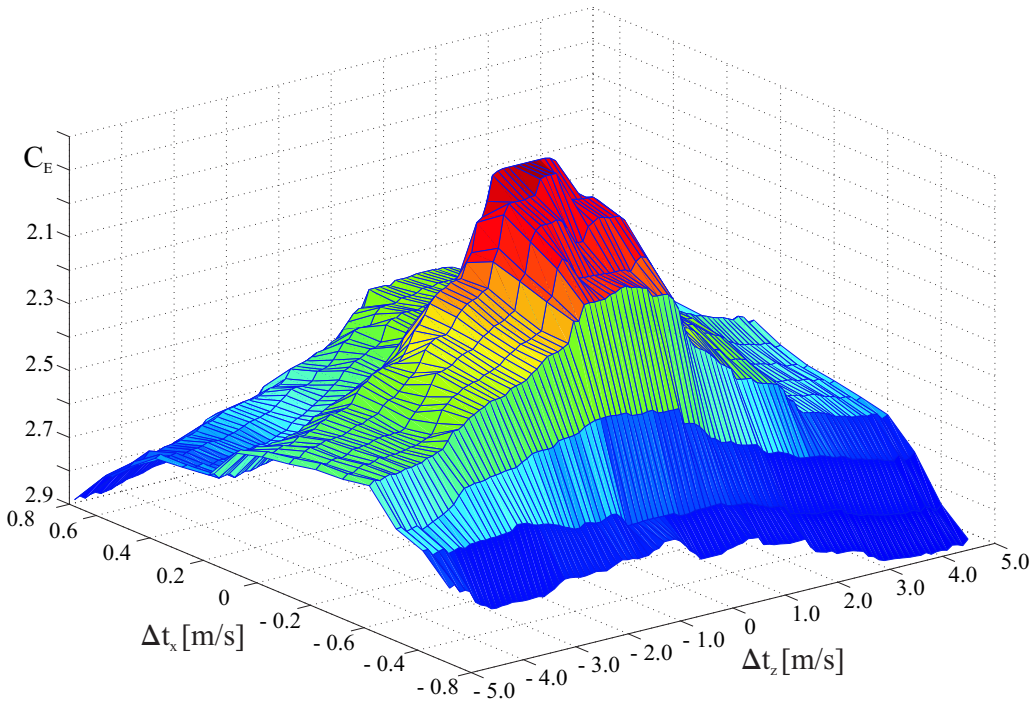
$\pm 0,5$ [m/s] und $\pm 0,2$[°/s] zunehmend deutlich vom Restfehler der optimalen Segmentierung abweicht. Auch hier kann die Begrenzung der Werte für $C_E$, trotz ansteigendem Bewegungsfehler, durch die zunehmende Aktivierung des Null-Labels erklärt werden.

## 5.3 Szenensegmentierung

Die Ergebnisse der Szenensegmentierung werden im Folgenden, entsprechend den beiden Teilschritten (i) Vorsegmentierung und (ii) bewegungsbasierter Objektsegmentierung, nacheinander vorgestellt.




(a)




(b)

**Abbildung 5.7**: (a) Restfehlerverteilung für zunehmenden Fehler der einzelnen Bewegungskomponenten. (b) Typische Restfehlerverteilung für den realistischen Fall mehrerer unsicherheitsbehafteter Bewegungskomponenten. Beispielhaft sind hier die beiden Bewegungskomponenten aufgeführt, die relativ gesehen einen großen Einfluss auf den Segmentierungsfehler haben.

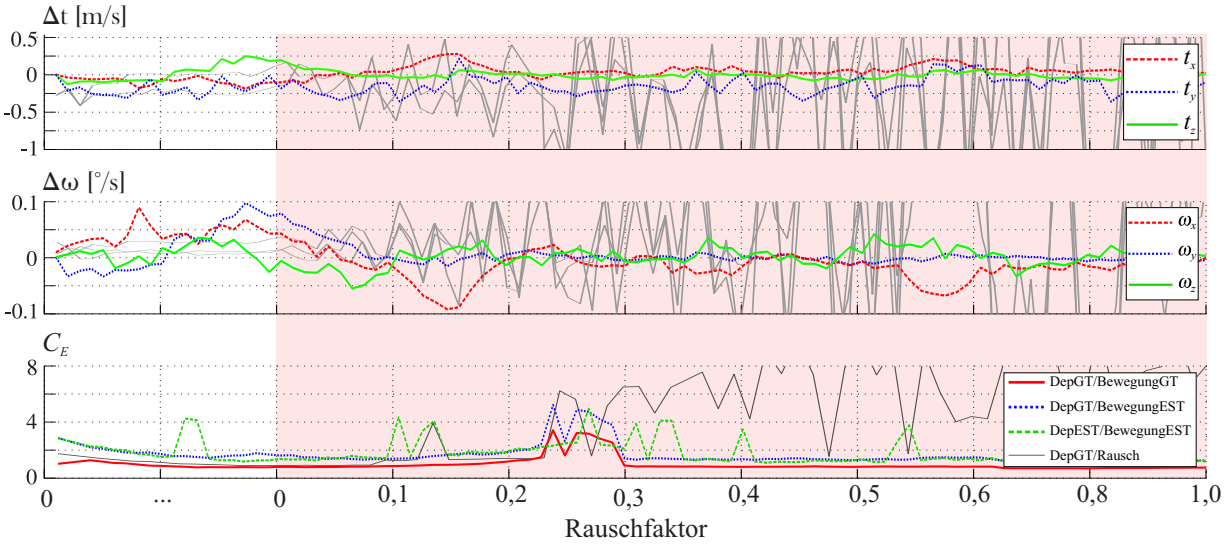



**Abbildung 5.8**: Einfluss der Schätzunsicherheit der Bewegung auf den Restfehler der Segmentierung für eine synthetisch generierte Bildsequenz mit bekannten Objektbewegungen und Tiefenwerten. Der obere Teil der Abbildung zeigt die Schätzfehler der translatorischen und rotatorischen Komponenten. Der Bereich, in dem der wahren Bewegung künstlich additives Rauschen überlagert wird, ist farblich markiert. Das Diagramm im unteren Teil zeigt den entsprechenden Restfehler $C_{\mathrm{E}}$ der Segmentierung. Bei der Auswertung des Restfehlers wurde die bekannte Tiefenkarte verwendet.

### (i) Vorsegmentierung

Das Verfahren zur Schätzung der Fahrbahnebene wurde hinsichtlich Robustheit gegenüber Ausreißern, Messunsicherheiten und Messausfällen untersucht. Bezüglich der Auswirkungen von Messunsicherheiten und Messausfällen auf den Schätzprozess, wurde verschiedenen synthetisch generierten Bildsequenzen ein additiver Rauschterm überlagert bzw. ein bestimmter, zufällig gewählter Anteil der Messdaten ausgeblendet. Die Untersuchung ergab, dass die Ebenenschätzung äußerst robust gegenüber Messausfällen ist. Bis zu einem verbleibenden Anteil von 20% an Messungen im gesamten Messbereich, waren die Auswirkungen auf das Schätzergebnis minimal. Eine Verschlechterung der Schätzung in Abhängigkeit des Messfehlers war erst ab einem Wert von $> 10\sigma_{\Delta}$ erkennbar. Große Auswirkungen auf die Ebenenschätzung haben jedoch systematisch im Messbereich auftretende Ausreißer. In der praktischen Anwendung werden solche Ausreißer typischerweise durch, in den Messbereich eintretende, Szenenobjekte hervorgerufen. Experimente mit künstlich generierten Objekten, die aus unterschiedlichen Richtungen in den Messbereich hereingeführt wurden, führten bereits bei einem

Ausreißeranteil von $< 20\%$ zu einer deutlichen Verschlechterung der Schätzergebnisse. Für den in der Praxis häufig auftretenden Fall einer Störung der Messungen durch ein frontal in den Messraum eintauchendes Objekt war der Effekt am Stärksten. Durch die zeitliche Glättung der Ebenenparameter mit einer angeschlossenen Plausibilitätsprüfung der aktuellen Schätzergebnisse konnte die Fahrbahnschätzung jedoch für eine praxistaugliche Anwendung nutzbar gemacht werden. Das Ergebnis der Ebenenschätzung für eine bildweise Auswertung und die in dieser Arbeit verwendete zusätzliche zeitliche Kopplung ist in Abbildung 5.9 grafisch dargestellt. Die Vorsegmentierung erlaubt bereits eine erste Einteilung der Szene in befahrbare Bereiche oder aber solche, die ein Hindernis darstellen. Bezüglich der Bewegungssegmentierung, die sich der Vorsegmentierung anschließt, konnte somit die Rechenzeit um bis zu 30% gegenüber der Auswertung des gesamten Bildinhalts reduziert werden.

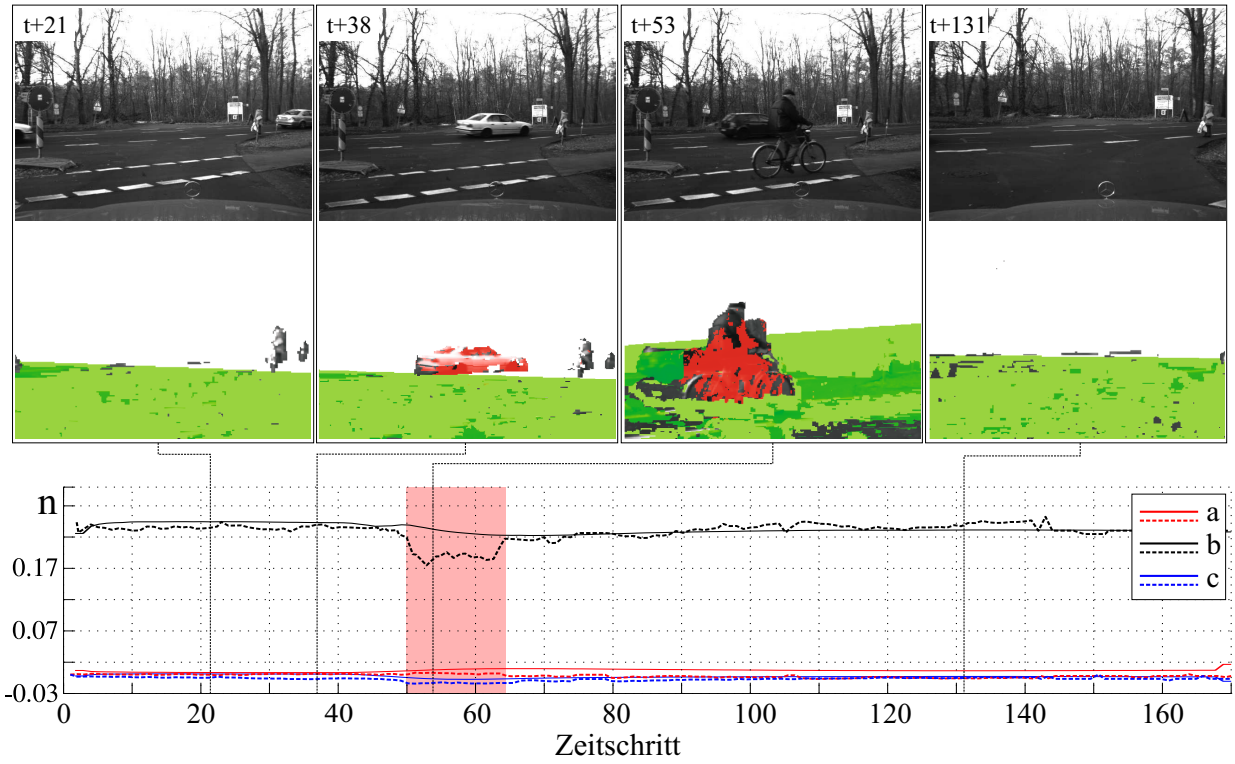



**Abbildung 5.9**: Oben: Ausschnitte einer Bildsequenz zu unterschiedlichen Zeitpunkten. Darunter ist das Segmentierergebnis gezeigt. Fremdbewegte Objekte sind rot marktiert. Die geschätzte Fahrbahnebene ist in grün eingeblendet. Unten: Die geschätzten Ebenenparameter $\mathbf{n} = (a, b, c)^{\mathrm{T}}$ über der Zeit. Die unterbrochenen Kurven zeigen hierbei das Ergebnis der bildweisen Auswertung, die durchgezogene Linien die zeitlich geglätteten Parameter. Zum Zeitpunkt $t \approx 50$ taucht ein Objekt in den Messbereich ein. Der entsprechende Bereich ist farblich markiert.

### (ii) Objektsegmentierung

Die Erzeugung von neuen Objekthypothesen beruht auf der Auswertung des Segmentierergebnisses zu jedem Zeitpunkt. Hierbei wird die Tatsache ausgenutzt, dass Bildpunkten, die nicht eindeutig einer initialisierten Objektinstanz zugewiesen werden können, das Null-Label aufgeprägt wird. Die durchgeführten Experimente haben ergeben, dass ab einer Differenzgeschwindigkeit von $\Delta\mathbf{v}_t \approx 2$ [m/s] fremdbewegte Gruppierungen von Merkmalspunkten verlässlich detektiert werden können. Die Bewegungsrichtung des jeweiligen Objekts und die augenblickliche Eigengeschwindigkeit der Kamera haben hier jedoch einen deutlichen Einfluss auf die Detektionsgüte. Abbildung 5.10 zeigt einige Segmentierergebnisse und den entsprechenden Restfehler über der Zeit aufgetragen. Die Segmentierung der rot umrandeten Bildsequenz erzeugt einen hohen Restfehler, der durch die feinen Bildstrukturen im oberen Teil der Szene erklärt werden kann. In diesen Bereichen ist eine eindeutige Zuweisung der korrespondierenden Bildpunkte in den räumlich

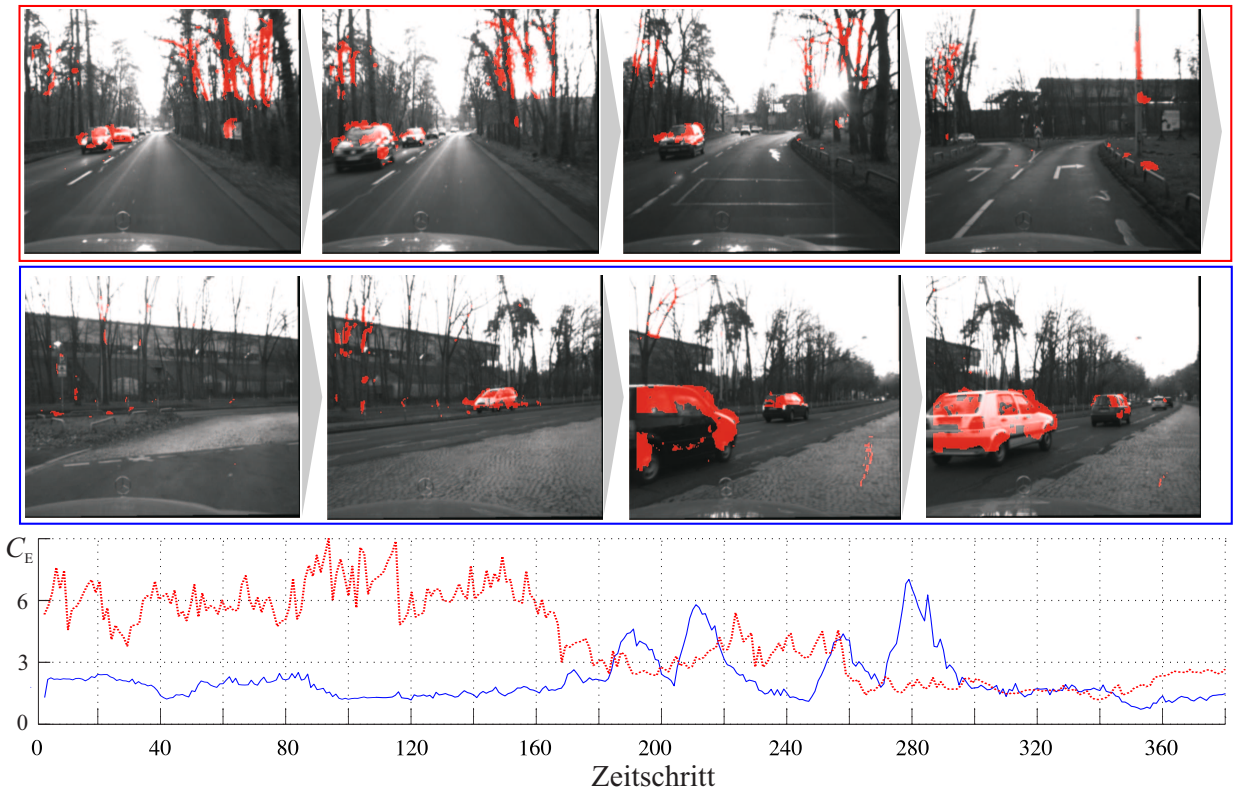



**Abbildung 5.10**: Oben: Momentaufnahmen einer Hintergrundsegmentierung. Unten: Restfehler der Segmentierung über der Zeit. Der Betrag des Restfehlers steigt deutlich an, falls größere Bildbereiche nicht mit der erwarteten Bewegung übereinstimmen. Die zugehörige Bildsequenz im oberen Teil ist in der Farbe der jeweiligen Kurve umrandet.

und zeitlich versetzten Ansichten der Szene nur bedingt möglich. Der Effekt kann teilweise durch eine höhere Bildauflösung oder die Anpassung der Parameter in der Vorsegmentierung reduziert werden.

Abbildung 5.11 zeigt den Segmentierungsfehler $C_L$ einzelner Bildbereiche über der Zeit für eine synthetisch generierte Bildsequenz. Der durch die fremdbewegten Objekte erzeugte Anstieg des Segmentierungsfehlers führt dazu, dass neue Objekthypothesen dem Segmentierprozess hinzugefügt werden. Wird die Initia-

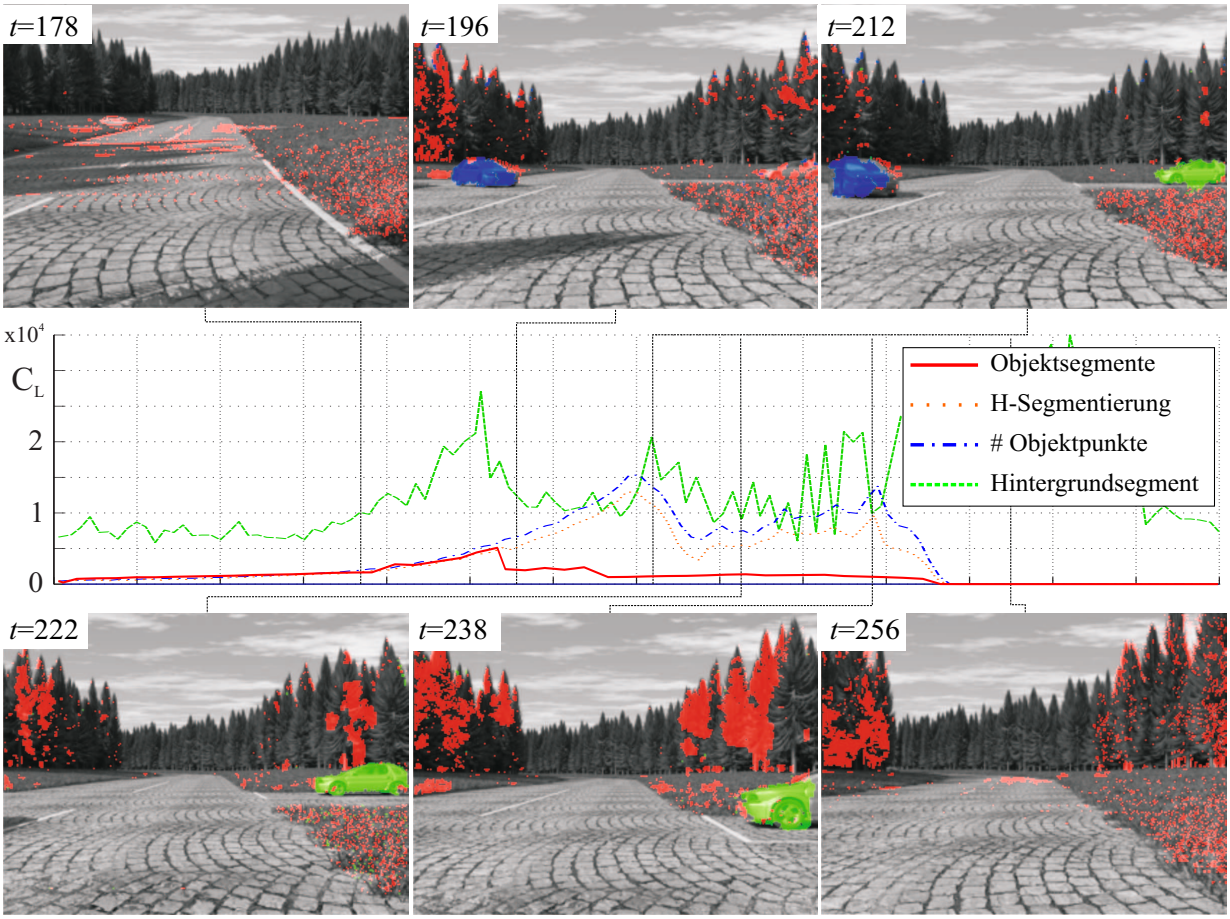



**Abbildung 5.11**: Die einzelnen Kurven zeigen den Fehler $C_L$ des Hintergrundsegments (– –) und der Objektsegmente. Die gepunktete Kurve (· · ·) zeigt hier den Fehler bei ausschließlicher Hintergrund(H-)segmentierung, in denen fremdbewegte Punkte dem Null-Label zugewiesen werden. Die durchgezogene Linie (——) zeigt den Standardfall, bei dem neue Objekthypothesen dem Prozess hinzugefügt werden. Die durch Punkte unterbrochene Linie (– · –) zeigt die absolute Anzahl fremdbewegter Punkte in der aktuellen Szene. Die Segmentierung ist den Bilddaten farblich überlagert. Rote Punkte zeigen das Null-Label, blau und grün symbolisiert die Labelwerte der jeweiligen Objekthypothese.

lisierung neuer Hypothesen unterdrückt, werden Punkte, die nicht der erwarteten Eigenbewegung der Kamera entsprechen dem Null-Label zugewiesen. Aufgrund der begrenzten Trennschärfe der einzelnen Objektbewegungen im Bild kann hierdurch jedoch nur ein Teil der Bildpunkte als fremdbewegt klassifiziert werden. Als Folge daraus wird ein Großteil der fremdbewegten Bildpunkte dem Hintergrund zugewiesen, wodurch der Fehler anwächst. In der Praxis hat sich jedoch gezeigt, dass ein solches Segmentierergebnis ausreicht, um fremdbewegte Bereiche im Bild verlässlich zu detektieren.

Die Laufzeit der einzelnen Module in Abhängigkeit der Anzahl initialisierter Objekthypothesen ist in Abbildung 5.12 aufgetragen. Das Verfahren skaliert annähernd linear mit der jeweiligen Menge an unabhängig bewegten Objekten in der Szene. Der größte Rechenaufwand entsteht durch die Rekonstruktion der Szene, die für den einfachsten Fall einer Hintergrundsegmentierung nahezu 70% des gesamten Rechenaufwands vereinnahmt. Mit zunehmender Anzahl an Objekthypothesen nimmt jedoch der Anteil der Bewegungsschätzung und der Bildsegmentierung zunehmend mehr Rechenzeit in Anspruch. Die Laufzeit der Ebenenschätzung mit anschließender Vorsegmentierung kann im Vergleich dazu als vernachlässigbar eingestuft werden.

In Abbildung 5.13 sind weitere Ergebnisse der Szenensegmentierung für eine qualitative Bewertung visualisiert. Als fremdbewegt klassifizierte Bildbereiche sind farblich unterlegt.

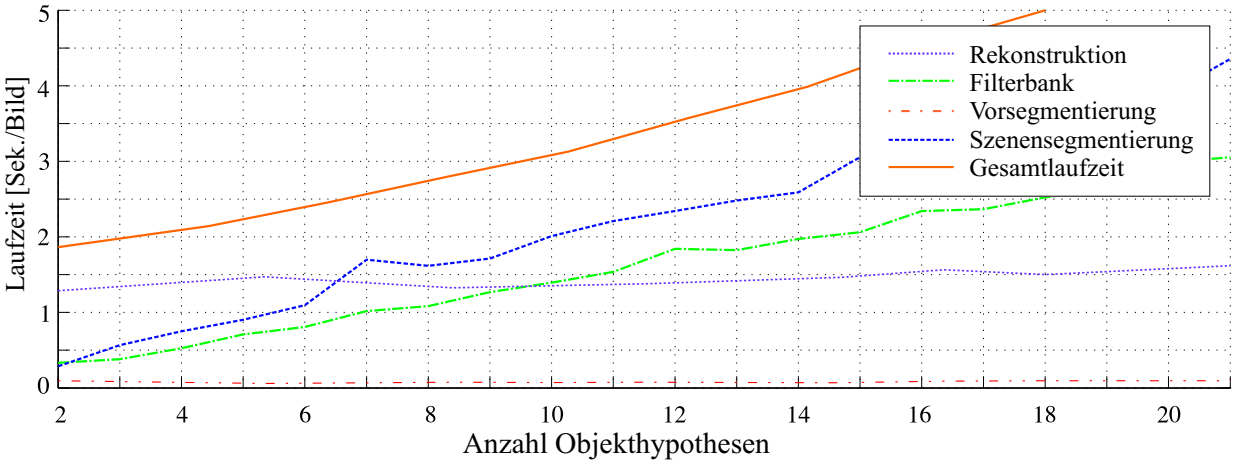



**Abbildung 5.12**: Laufzeit der einzelnen Module des Verfahrens und die Gesamtlaufzeit in Abhängigkeit der Anzahl initialisierter Objekthypothesen. Die dargestellten Rechenzeiten ergaben sich bei einer Bildauflösung von 512x384 Bildpunkten auf einem Rechner mit Intel Core 2 Prozessor (2,8GHz, 2GB RAM).

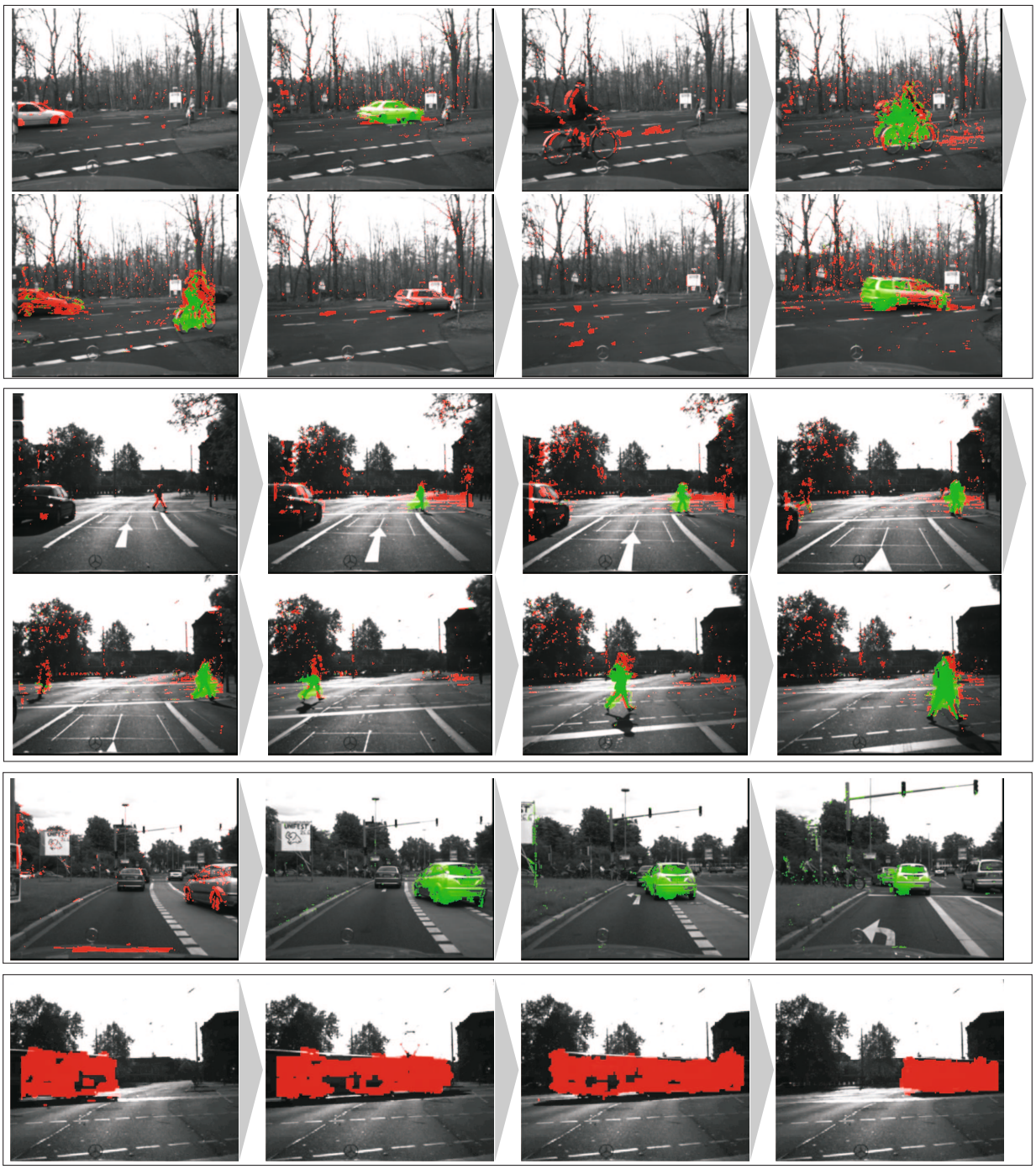


**Abbildung 5.13**: Ergebnisse der Szenensegmentierung für eine Auswahl von Bildsequenzen, die alltägliche Verkehrsszenarien zeigen. Als fremdbewegt klassifizierte Bildpunkte sind rot unterlegt. Bereiche, die einer unabhängig bewegten Objektinstanz zugewiesen werden, sind farblich grün markiert. Das untere Bild zeigt das Segmentierergebnis für den Fall sehr starker Glättung. Das detektierte Objekt ist hier rot markiert.

KAPITEL 6

# Zusammenfassung und Ausblick

In dieser Arbeit wird ein Verfahren zur Segmentierung unabhängig bewegter Verkehrsobjekte in Stereobildfolgen vorgestellt. Besonderer Wert wird hierbei auf eine möglichst vollständige und für den Menschen unmittelbar interpretierbare Zerlegung des Bildes in physikalisch relevante Szenenbereiche gelegt. Hierfür wird die Szene parametrisch durch eine Menge sich starr im dreidimensionalen Raum bewegender Objekte beschrieben.

Bezüglich der statischen Szene wird zusätzlich die Annahme einer ebenen Welt getroffen. Unter dieser Annahme können in einem Vorsegmentierungsschritt geometrisch nicht plausible Szenenbereiche bereits zu einem sehr frühen Zeitpunkt aus der Verarbeitungskette herausgenommen werden.

Ausgehend von einer Bayes'schen Formulierung der Szenensegmentierung für die verbleibenden Bereiche im Bild, wird ein Maximum-A-Posteriori-Gütemaß vorgestellt, welches gleichermaßen die Segmentierung und die Schätzung der Parameter des Szenenmodells bewertet. Die einzelnen Modellparameter der Objekthypothesen werden auf der Basis eindeutig bestimmbarer Bildmerkmale geschätzt, wodurch eine hohe Robustheit des Verfahrens erreicht wird. Für Objekte, die relativ zur Kamera fremdbewegt sind, wird das Modell um die Eigenschaft endlicher räumlicher Ausdehnung erweitert und in den Segmentierprozess integriert. Um eine Segmentierung an jeder Bildposition durchführen zu können, wird die Szene vollständig stereoskopisch rekonstruiert, wobei zusätzlich die zeitliche Entwicklung der Tiefenwerte berücksichtigt wird. Hierbei wird angenommen, dass sich ein unabhängig bewegtes Szenenobjekt entlang seiner geschätzten Bewegungstrajektorie bewegt, womit gewisse Konsistenzforderungen an die Szenenrekonstruktion und die Segmentierung gestellt werden können. Der intuitiven Erwartung stückweise stetiger und gleichförmig bewegter Objektoberflächen wird bei der Beschreibung des Szenenmodells explizit durch das Modell eines Markov-Zufallsfeldes Rechnung getragen.

Zur Lösung des verkoppelten Optimierungsproblems der dichten Szenensegmentierung wird die Schätzung der Objektparameter in alternierender Reihenfolge, ähnlich dem bekannten Expectation-Maximization-Verfahren, gelöst. Dabei werden die Beobachtungen nicht entsprechend einer binären Entscheidungsregel nur einer bestimmten Objekthypothese zugewiesen, sondern probabilistisch gewichtet auf die Menge aller Objekthypothesen abgebildet. Glattheitsanforderungen an den Assoziationsprozess werden durch das gleiche Modell wie bei der Segmentierung beschrieben. Zur Schätzung der Bewegungsparameter wird ein Kalman-Glätter vorgestellt, welcher in iterativer Weise das Schätzergebnis verbessert. Dabei wechselt das Verfahren zwischen einer Filterung in Vorwärtsrichtung durch die Zeit und einer Glättung der Schätzergebnisse in Rückwärtsrichtung. Die Glättung wird dabei auf einen für die Anwendung im Straßenverkehr sinnvollen Bereich beschränkt. Durch die Kopplung der merkmalsbasierten, rekursiv durchgeführten Parameterschätzung mit der dichten Rekonstruktion und Segmentierung der Szene, können zudem die zeitlichen Abhängigkeiten der Daten für das gesamte Bild nutzbar gemacht werden.

Die Bestimmung der im Schätzprozess enthaltenen Hypothesenmenge erfolgt durch ein heuristisch motiviertes Ballungsverfahren, welches auf der Grundlage des aktuell vorliegenden Segmentierergebnisses eine Konsistenzprüfung aller, als fremdbewegt detektierten Merkmalspunkte, durchführt. Somit können effizient neue Objekthypothesen dem Prozess hinzugefügt werden. Die Vernichtung unplausibler Hypothesen erfolgt durch die Segmentierung selbst.

Die Leistungsfähigkeit des Verfahrens wird auf der Basis realer und synthetisch generierter Stereobildsequenzen untersucht und bewertet. Neben einer objektiven Bewertung der Qualität des Verfahrens anhand von aussagekräftigen Gütemaßen, wird der Ansatz durch zahlreiche visualisierte Segmentierergebnisse auch einer subjektiven Bewertung zugänglich gemacht.

Ein Alleinstellungsmerkmal dieser Arbeit ist die ganzheitliche Beschreibung der Segmentieraufgabe, welche die Teilaufgaben Szenenrekonstruktion, Bewegungsschätzung und Segmentierung in einem Modell zusammenfasst. Hierdurch können die starken gegenseitigen Bindungen der einzelnen Größen für die Schätzaufgabe nutzbar gemacht und in eine gemeinsame Optimierungsstruktur eingebettet werden. Die modulare Systemarchitektur erlaubt dabei den einfachen und komfortablen Austausch einzelner Komponenten. Das probabilistische Modell zur Szenensegmentierung ist weiterhin sehr flexibel bezüglich der Integration beliebig abstrahierter Szeneninformation. Die Tatsache, dass diese Information bei der Schätzung der Modellparameter durch den Datenassoziationsprozess mit berücksichtigt wird, unterstreicht nochmals das hohe Potenzial des Verfahrens.

Wie in Kapitel 5 gezeigt, liefert das hier vorgestellte Verfahren zur Szenensegmentierung für viele Verkehrsszenarien bereits gute Ergebnisse. Jedoch sind noch einige Verbesserungen denkbar. So erscheint bzgl. der Implementierung des Verfahrens eine Parallelisierung einzelner Algorithmen als sinnvoll. Speziell die rechenaufwändige dichte Rekonstruktion der Szene und die Bildsegmentierung bieten diesbezüglich großes Optimierungspotential. Das Verfahren sollte weiter in unterschiedlichen Verkehrsszenarien unter verschiedenen Witterungsbedingungen getestet werden. Erste Untersuchungen hierzu haben gezeigt, dass der Einfluss von Schnee und Regen zu einer deutlichen Verschlechterung der Ergebnisse führt. Im Rahmen dieser Arbeit wurde weiterhin deutlich, dass die Szenensegmentierung sehr eng mit den parallel am Institut für Mess- und Regelungstechnik durchgeführten Arbeiten zur dichten Schätzung des optischen Flussfeldes sowie der dreidimensionalen Registrierung verknüpft ist. Durch die Kombination der jeweiligen Verfahren mit der Szenensegmentierung können hier sicherlich Synergieeffekte auf beiden Seiten erwartet werden.

Mit dem Ziel einer vollständigen Szeneninterpretation, entsprechend dem hierarchischen Modell aus Abbildung 1.1, besteht der nächste logische Schritt aus einer Klassifikation der Szenenobjekte. Das dadurch gewonnenen „Verständnis" räumlicher und semantischer Beziehungen der in der Szene befindlichen Objekte zueinander, kann schließlich eine Maschine zu einer „intelligenten" Interaktion mit ihrer Umwelt befähigen. Für die zukünftige Entwicklung sicherheitsrelevanter Systeme erscheint diese Form der Umgebungswahrnehmung zwingend notwendig, um die heute meist *re*-agierenden Systeme zu agierenden Systemen weiterzuentwickeln, die ihre Handlung der jeweiligen Situation anpassen. Die dichte Objektsegmentierung in Bildfolgen erscheint hierfür besonders geeignet, da der Informationsgehalt der Bilddaten nicht schon auf den unteren Ebenen der Szeneninterpretation durch stark vereinfachende Modellannahmen ersetzt wird. Erste Versuche zur Integration objektspezifischen Wissens wurden im Rahmen dieser Arbeit bereits erfolgreich durchgeführt und zur Verbesserung der Segmentierung genutzt.

Das Potenzial und die Umsetzbarkeit einer bildbasierten dichte Objektsegmentierung wurden aufgezeigt. Für die weitere Entwicklung im Bereich der mobilen Umfeldwahrnehmung wird diese Form der Szenenrepräsentation sicher eine wichtige Rolle spielen und ihren Beitrag dazu leisten, ein neues Verständnis für die menschliche Mobilität zu prägen.

# Anhang

## A.1 Globale Optimierungsverfahren

Allgemein zeichnen sich Verfahren zur Lösung eines Ausdrucks wie in (3.29) bzw. (3.30) beschrieben dadurch aus, dass sie entweder ein *lokales* oder aber ein *globales* Minimum des Energiefunktionals $H(\mathbf{l})$ finden. Die Lokalität, bzw. Globalität eines Minimums $\mathbf{l}$ ist dabei in Bezug auf ein Nachbarschaftssytem $\mathcal{N} = \{\mathcal{N}_\mathbf{l} \,|\, \mathbf{l} \in \mathcal{F}\}$ definiert. $\mathcal{N}_\mathbf{l}$ drückt die Menge der Nachbarschaftskonfigurationen von $\mathbf{l}$ im Lösungsraum aus, wobei $\mathcal{F}$ den Konfigurationsraum von $\mathbf{l}$ darstellt. Die Nachbarschaft sei definiert durch $\mathcal{N}_\mathbf{l} = \{\mathbf{f} \,|\, \mathbf{f} \in \mathcal{F}, \|\mathbf{f}-\mathbf{l}\| > 0\}$, wobei $\|\cdot\|$ die absolute Differenz der beiden Labelrealisationen in $\mathbf{f}$ und $\mathbf{l}$ ausdrückt. Eine Konfiguration $\mathbf{l}^*$ erzeugt nun ein lokales Minimum relativ zur Nachbarschaft $\mathcal{N}$, falls

$$H(\mathbf{l}^*) \leq H(\mathbf{l}) \quad \forall \mathbf{l} \in \mathcal{N}_{\mathbf{l}^*} \,. \tag{A.1}$$

Es geht per Definition in ein globales Minimum über, falls $\mathcal{N}_\mathbf{l} = \{\mathbf{l} \,|\, \forall \mathbf{l} \in \mathcal{F}, \mathbf{l} \neq \mathbf{l}^*\} = \mathcal{F} - \{\mathbf{l}\}$. Die jeweiligen Verfahren zum Auffinden eines lokalen oder globalen Minimums des Energiefunktionals suchen nun, ausgehend von einer Konfiguration $\mathbf{l}$, nach einer Alternativkonfiguration $\mathbf{l}'$ in der Nachbarschaft $\mathcal{N}_\mathbf{l}$. Hat die Veränderung zur Folge, dass $H(\mathbf{l}') < H(\mathbf{l})$, wird $\mathbf{l}$ durch $\mathbf{l}'$ ersetzt. Dieser Prozess wird solange wiederholt, bis keine Verbesserung mehr eintritt.

### A.1.1 Verfahren aus der Literatur

#### Lokale Verfahren

Von [Besag, 1986] wird ein deterministisches Verfahren vorgestellt, welches iterativ die jeweilige lokale Mode in Abhängigkeit seiner Nachbarschaft bestimmt

und deshalb als *iterated conditional modes* (ICM) bezeichnet wird. Für eine gegebene Bildsequenz $\mathcal{G}_T$ und Labeling $\mathbf{l}^k_{\mathcal{I}_{\setminus n}}$, berechnet das Verfahren iterativ jede Labelvariable $\mathrm{l}_n^{k+1}$ durch die lokale Maximierung der bedingten Wahrscheinlichkeit $P(\mathrm{l}_n|\mathcal{G}_T, \mathbf{l}_{\mathcal{I}_{\setminus n}})$. Neben der Bedingung stochastischer Unabhängigkeit der Beobachtungen wird bei der Auswertung dieses Ausdrucks angenommen, dass $\mathbf{l}$ nur von seinen lokalen Nachbarn abhängig ist. Unter Verwendung des Bayes-Theorems folgt hieraus

$$P(\mathrm{l}_n|\mathcal{G}_T, \mathbf{l}_{\mathcal{I}_{\setminus n}}) \propto P(\boldsymbol{\varepsilon}_{n,t}|\mathrm{l}_n, \hat{\mathbf{l}}_{t-1}, \boldsymbol{\Theta}_t) P(\mathrm{l}_n|\mathrm{l}_{\mathcal{N}_n}) \,. \tag{A.2}$$

Beim ICM-Verfahren wird obige Gleichung für jede Labelvariable $\mathrm{l}_n$ , $n \in \mathcal{I}$ einzeln ausgewertet. Nach der lokalen Optimierung einer Variablen setzt das Verfahren an einer zufällig gewählten Stelle innerhalb des Labelfeldes fort. Dadurch sollen unerwünschte systematische Verfälschungen des Ergebnisses verhindert werden.

In [Chou u. Brown, 1990] wird ein, dem ICM-Verfahren sehr ähnlicher Ansatz vorgestellt, der als *Highest-Confidence-First*-Verfahren (HCF) bezeichnet wird. Dabei werden die einzelnen Variablen des Zufallsfeldes nicht in zufälliger Reihenfolge abgearbeitet, sondern es werden Punkte bevorzugt, die eine signifikante Auswirkung auf das Energiefunktional aufweisen. Hierfür wird die Labeldefinition um einen zusätzlichen Wert erweitert, der die „Stabilität" der lokalen Labelzuweisung codiert. Somit kann der Optimierungsprozess auf Bereiche konzentriert werden, die noch nicht hinreichend optimiert sind, bzw. Punkte mit einem hohen Informationsgehalt werden bevorzugt optimiert. Im Vergleich mit ICM kann durch diese Strategie der Rechenaufwand deutlich reduziert werden.

Wie bei allen lokalen Optimierungsverfahren, besteht bei den oben aufgeführten Verfahren die Gefahr in ein lokales Minimum des Energiefunktionals zu fallen und dort zu verharren. Um dieses Problem zu umgehen, können globale Verfahren eingesetzt werden, die teilweise an die lokalen Ansätze anknüpfen.

### Globale Verfahren

Um das Problem mit den Nebenminima bei lokalen Verfahren zu umgehen, gibt es eine Reihe von Ansätzen, die im Folgenden kurz vorgestellt werden. So wird beim *Simulierten Abkühlen* (engl. *simulated annealing*) der physikalische Abkühlvorgang einer Kristallstruktur nachgebildet. Anstatt eines Gradientenabstiegs wird hier ein randomisiertes Suchverfahren zur Bestimmung einer neuen Konfiguration eingesetzt. Als Suchverfahren kann der bekannte Gibbs-Sampler [Geman u. Geman, 1984] oder der Metropolis-Hasting Algo-

rithmus [Metropolis u. a., 1953] verwendet werden. In ihrer Funktionsweise können Sie als Zufallssuchalgorithmen verstanden werden, die auch Labelkonfigurationen erlauben, die die Energie kurzzeitig anwachsen lassen. Gerade durch dieses kurzzeitige Anwachsen können lokale Minima überwunden werden. Die Suche wird durch einen Kontrollparameter $T$ gesteuert. Analog zum physikalischen Modell des Abkühlvorgangs in einer Kristallstruktur, erlauben hohe Werte von $T$ einen großen Anstieg der Energie. Für kleiner werdendes $T$ reduziert sich auch die Akzeptanz für ein Anwachsen des Energieterms bis hin zum einfrieren des Systems, bei dem nur noch abnehmende Energiewerte akzeptier werden.

Als deterministische Verfahren sind die Methode der *graduiert relaxierten Konvexität* (engl. *graduated nonconvexity*) [Blake u. Zisserman, 1987] und die *Mean-Field-Annealing*-Methode [Yuille, 1987] zu nennen.

### A.1.2 Binäres Graphenschnittverfahren

Der Max-Flow/Min-Cut-Algorithmus von Boykov und Kolmogorov [Boykov u. Kolmogorov, 2001] gehört zur Gruppe der flusserhöhenden Pfadsuchverfahren. Der wesentliche Unterschied zu den bekannten Verfahren in dieser Gruppe besteht darin, dass es zwei Suchbäume Q und S gibt, die zeitgleich aufeinander zuwachsen. Die Wurzeln der Bäume sind jeweils die Quelle $q$ und die Senke $s$. Die Suchbäume sind disjunkt, bilden aber keine vollständige Zerlegung des Graphen $\mathrm{G}(\mathcal{P},\mathcal{E})$. Für den Baum Q gilt, dass jede Kante vom Elternknoten zum Kindknoten nicht gesättigt ist. Im Baum S gilt das Gleiche für Kanten von den Kindknoten zu den Elternknoten. Die Knoten der Suchbäume können zwei Zustände annehmen: *aktiv* (Knoten mit nicht erfassten Nachbarn) und *passiv* (innere Knoten). Aktive Knoten besitzen die Eigenschaft, einen Suchbaum zu erweitern. Trifft ein aktiver Knoten auf einen Nachbarn des anderen Suchbaumes, ist ein flusserhöhender Pfad gefunden. Der Algorithmus lässt sich in folgende drei Phasen aufteilen:

- **Wachstumsphase**
  In dieser Phase findet eine Erweiterung von Q und S statt, bis ein erster geschlossener $\langle q,s\rangle$-Pfad gebildet wird. Hierfür akquirieren aktive Knoten die über ungesättigte Kanten verbundenen Nachbarknoten welche noch keinem Suchbaum angehören. Entsprechend werden die akquirierten Nachfolgeknoten als aktiv markiert. Der Status eines aktiven Knotens ändert sich, falls alle seine Nachbarknoten untersucht worden sind. Die Wachstumsphase terminiert, sobald ein aktiver Knoten einen Nachbarknoten untersucht der zum anderen Suchbaum gehört.

◇ **Verstärkungsphase**
Der Fluss über den in der Wachstumsphase gefundenen Pfad wird nun maximal erhöht. Dadurch erhält man mindestens eine gesättigte Kante im Pfad. Anschließend werden die Kanten in Flussrichtung $q \rightarrow s$ des Pfades um diese Kapazität verringert, während die Kanten in entgegengesetzter Richtung $s \rightarrow q$ um diese Kapazität erhöht werden. Es ist nun möglich, dass eine Kante nach der Flusserhöhung gesättigt ist, wodurch die Suchbäume in einzelne Teilbäume zerfallen. Die beiden beteiligten Knoten werden als *Eltern*- und *Waisenknoten* (engl. *orphans*) bezeichnet. Sie bilden die Wurzelknoten der neuen Teilbäume. Damit die Suchbaumstruktur mit jeweils der Quelle und der Senke als Wurzel wieder hergestellt werden kann, ist im dritten Schritt eine Adoptions- oder Anpassungsphase nötig.

◇ **Anpassungsphase**
In dieser Phase werden die ursprünglichen Bäume Q und S wiederhergestellt, d. h. es wird versucht, für jeden Waisenknoten einen gültigen Elternknoten zu finden. Als Anpassungsbedingung gilt die Forderung, dass Eltern- und Waisenknoten zum selben Suchbaum gehören müssen. Zusätzlich müssen der Eltern- und Waisenknoten über eine ungesättigte Kante verbunden sein. Ist kein gültiger Elternknoten vorhanden, wird der Waisenknoten aus dem Suchbaum entfernt und als freier Knoten markiert. Als Folge daraus werden alle direkten Nachfolgeknoten des nun freien Knotens zu Waisenknoten. Die Anpassungsphase endet mit der Abarbeitung des letzten Waisenknotens. Als Resultat der Anpassungsphase erhält man eine Suchbaumstruktur mit den beiden Bäumen Q und S, womit der Prozess wieder in die Wachstumsphase übergehen kann.

Abbildung A.1 verdeutlicht die Arbeitsweise des Verfahrens grafisch. Die Abfolge der drei Phasen wird solange wiederholt, bis der maximale Fluss gefunden ist,

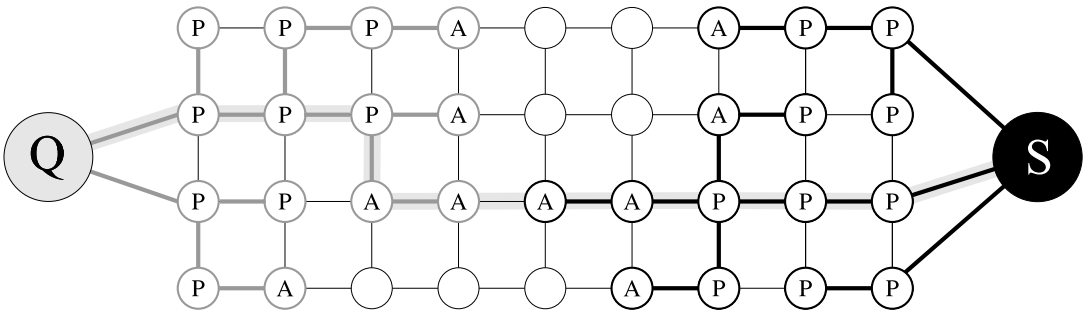



**Abbildung A.1**: Die Wurzeln der Suchbäume Q und S stellen die Abschlussknoten des Graphen dar. Aktive Knoten sind mit A, passive Knoten mit P beschriftet.

d. h. bis die Suchbäume Q und S keine aktiven Knoten mehr enthalten. Eine ausführliche Beschreibung findet sich in [Kolmogorov u. Zabih, 2004].

### A.1.3 Der $\alpha$-Expansion-Algorithmus

Durch die hier vorgestellte iterative Optimierungsstrategie kann eine suboptimale Zerlegung des Graphen in $J > 2$ Klassen erreicht werden, wie in Abbildung A.2 skizziert. In jedem Iterationsschritt wird ein Teil des Gesamtproblems auf ein binäres Optimierungsproblem abgebildet, welches dann z. B. mit dem in Anhang A.1.2 beschriebenen Verfahren gelöst werden kann. Der Algorithmus arbeitet in iterativer Weise nach folgendem Prinzip:

Es wird eine initiale Startkonfiguration $\mathbf{l} = \mathbf{l}^0$ vorgegeben. In jedem Iterationsschritt wird nun ein Label $\alpha$ fest gewählt. Die Komponenten im Labelfeld $\mathbf{l}$, denen bereits der Labelwert $\alpha$ aufgeprägt ist, werden fixiert. Für alle anderen Komponenten $l_n \neq \alpha$ muss entschieden werden, ob der aktuelle Wert auf $\alpha$ gesetzt werden soll, d. h. eine Erweiterung (engl. *expansion*) der Klasse $\alpha$ um die entsprechende Komponente im Labelfeld stattfinden soll. Von grundlegender Bedeutung sind hier die sog. zulässigen Schritte im Konfigurationsraum $\mathcal{F}$.

**Definition A.1 (Schritt)**
*Ein zulässiger Schritt im Konfigurationsraum $\mathcal{F}$ ist ein Paar $(\mathbf{l}^1, \mathbf{l}^2)$ mit $\mathbf{l}^1, \mathbf{l}^2 \in \mathcal{F}$. Es handelt sich um einen zulässigen $\alpha$-Erweiterungsschritt, falls*

$$\forall n \in \{1, \ldots, N\} : (l_n^2 = l_n^1) \cup (l_n^2 = \alpha)\,. \tag{A.3}$$

*Die Menge des durch diese zulässigen Schritte erzeugten Labelfeldes $\mathbf{l}^2$, definiert die Umgebung $U^\alpha(\mathbf{l}^1)$.*

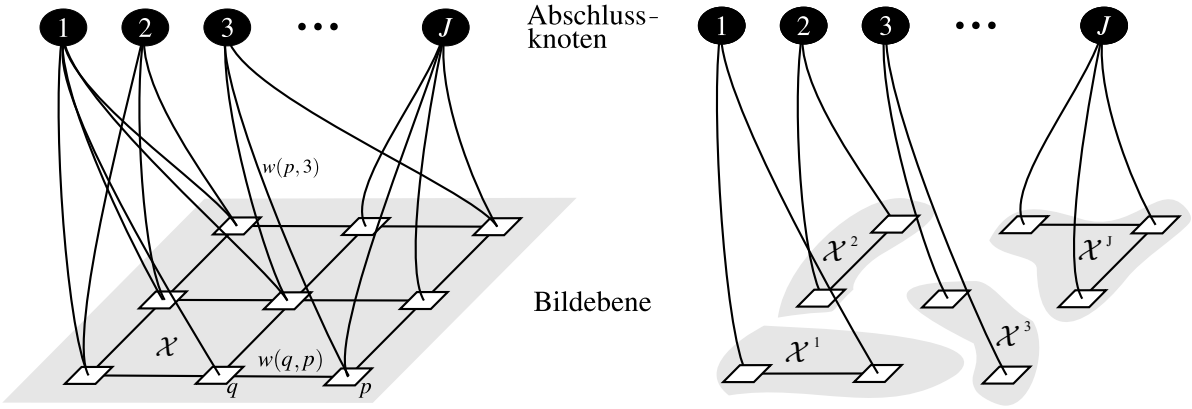



**Abbildung A.2**: Multiway-Graphenschnitt eines Bildes in $J$ Klassen.

Zu jedem Iterationsschritt $k$ erfolgt die Auswahl eines fixen Labels $\alpha^k$. Ausgehend von dieser Wahl wird dann das folgende allgemeine Teilproblem gelöst

$$\hat{\mathbf{l}}^{k+1} = \underset{\mathbf{l}\in U^{\alpha^k}(\mathbf{l}^k)}{\arg\min} \{F(\mathbf{l})\}\,, \tag{A.4}$$

wobei $F$ eine graphrepräsentierbare Funktion definiert.

Durch die Wahl der Umgebung $U$ kann das Problem als binäres Problem beschrieben werden, wie im Folgenden gezeigt wird. Zunächst ist durch die aktuelle Labelkonfiguration $\mathbf{l}^k$ und die Wahl des Labels $\alpha^k$ die Energiefunktion in der Umgebung nur noch von einem binären Feld $B$ abhängig, welches kodiert, ob eine Variable seinen Wert beibehält oder auf $\alpha$ gesetzt wird. Somit kann eine neue, binäre Energiefunktion definiert werden

$$F_b(B) = F(\mathbf{l}(B))\,,\ \text{ mit }\ \mathrm{l}_n(B) = \begin{cases} \alpha(k) & B_n = 1 \\ \mathrm{l}_n^k & B_n = 0\,. \end{cases} \tag{A.5}$$

Die Variablen im Labelfeld, denen bereits der Wert $\alpha$ aufgeprägt war, werden fixiert durch die zusätzliche Einschränkung $\forall n\, \mathrm{l}_n^k = \alpha(k) : B_n = 1$. Hieraus ergibt sich folgendes binäre Teilproblem:

$$\hat{\mathbf{l}}^{k+1} = \Gamma(\mathbf{l}^k,\ \underset{\substack{B\in\{0,1\}^N \\ \forall n\, \mathrm{l}_n^k=\alpha^k : B_n=1}}{\arg\min}\ \{F_b(B)\})\,. \tag{A.6}$$

Die Funktion $\Gamma$ dient dazu, die Information des binären Vektors $B$ zu dekodieren und ist durch (A.5) definiert, d. h. $\Gamma(\mathbf{l}^k, B) = \mathbf{l}(B)$. Das in obige Gleichung eingebettete binäre Optimierungsproblem entspricht der Beschreibung aus (2.35) und kann somit auf das Problem des minimalen Schnittes in einem Graphen reduziert werden. Durch die Fixierung der Labelwerte kann das Problem weiterhin vereinfacht werden zu

$$\underset{\substack{B\in\{0,1\}^N \\ \forall n\, \mathrm{l}_n^k=\alpha^k : B_n=1}}{\arg\min}\ \{F_b(B)\} = \underset{\tilde{B}\in\{0,1\}^{N-|\omega|}}{\arg\min}\ \left\{F_b(\tilde{B})\right\}\,, \tag{A.7}$$

wobei die Indexmenge $\omega = \{n | \mathrm{l}_n^k = \alpha^k\}$ hier alle fixierten Komponenten des Labelfeldes umfasst.

Um das binäre Optimierungsproblem in (A.7) mit Graphenschnittverfahren zu lösen, müssen nach Theorem 2.1 gewisse Anforderungen an $F_b$ erfüllt sein. Nimmt man für $F_b$ die allgemeine Form (2.36) an, kann gezeigt werden, dass sich die Bedingung aus (2.37) in folgender Weise auf das oben dargestellte Problem überträgt [Boykov u. a., 2001; Kolmogorov u. Zabih, 2004; Rodner, 2007]:

$$F^{n,u}(\mathrm{l}_n^k, \mathrm{l}_u^k) + F^{n,u}(\alpha,\alpha) \leq F^{n,u}(\mathrm{l}_n^k, \alpha) + F^{n,u}(\mathrm{l}_u^k, \alpha) \forall n, u : \mathrm{l}_n^k \neq \alpha \neq \mathrm{l}_u^k\,. \tag{A.8}$$

Ist diese Forderung erfüllt, lässt sich die Funktion $F$ mit dem $\alpha$-expansion-Verfahren minimieren. Für das in dieser Arbeit verwendete erweiterte Potts-Modell können die gestellten Bedingungen als erfüllt betrachtet werden, womit das Energiefunktional zur Schätzung der Szenensegmentierung graphrepräsentierbar ist und mit Hilfe des $\alpha$-expansion-Algorithmus effizient gelöst werden kann.

## A.2 Das Ebenenmodell

Im globalen Koordinatensystem, welches in dieser Arbeit mit dem Kamerakoordinatensystem der rechten Kamera zusammenfällt, wird die Ebene beschrieben durch

$$\mathbf{n}^{\mathrm{T}}\mathbf{X} = h, \tag{A.9}$$

mit Ebenenpunkt $\mathbf{X}$ in kartesischen Koordinaten und Parametervektor $\mathbf{n} = (a,b,c)^{\mathrm{T}}$. $h$ entspricht der Distanz vom Koordinatenursprung der Kamera zur Ebene und wird im weiteren Verlauf dieser Arbeit auf $h = 1$ gesetzt.

Mit Hilfe einer vollständig kalibrierten Stereokamera und der Beziehung aus (2.17) kann nun (A.9) in den Bildbereich transformiert werden. Durch die Transformation in den Bildbereich kann die hohe Fehlerempfindlichkeit bei der Ebenenschätzung mit zunehmender Szenentiefe verkleinert werden. Für die Ebenenschätzung wird ein Total-Least-Squares-Ansatz gewählt, der die orthogonalen Abstände zwischen gemessenem Ebenenpunkt und korrespondierendem Modellpunkt minimiert. Unter Verwendung von (2.17) kann mit der Normierungskonstante $\varphi$ ein Bildpunkt der Ebene $\mathbf{x}_e$ im Disparitätsraum wie folgt geschrieben werden:

$$\frac{\mathrm{b}}{\varphi} a x_e + \frac{\mathrm{b}}{\varphi} b y_e - \Delta_e = -c\mathrm{f}\,. \tag{A.10}$$

Anhand dieser Gleichung ist ersichtlich, dass die idealen und auf die Varianz normierten Beobachtungen

$$\mathbf{y'}_e = \begin{pmatrix} \frac{x_e}{\sigma_x^2} \\ \frac{y_e}{\sigma_y^2} \\ \frac{\Delta_e}{\sigma_\Delta^2} \end{pmatrix} = \begin{pmatrix} x'_e \\ y'_e \\ \Delta'_e \end{pmatrix} \tag{A.11}$$

für jeden gegebenen Parametersatz $(\mathbf{n}, \varphi)$ auf der Ebene $\mathcal{P}$ liegen. Wird $\mathbf{y'}_e$ in (A.10) eingesetzt, ergibt sich

$$\underbrace{\frac{\mathrm{b}}{\varphi} a \sigma_x^2}_{\mathbf{r}_1} x'_e + \underbrace{\frac{\mathrm{b}}{\varphi} b \sigma_y^2}_{\mathbf{r}_2} y'_e + \underbrace{\left(-\sigma_\Delta^2\right)}_{\mathbf{r}_3} \Delta'_e = -c\mathrm{f}\,, \quad \text{bzw. } \mathbf{r}^{\mathrm{T}}\mathbf{y'}_e = -c\mathrm{f}\,. \tag{A.12}$$

Für jeden beliebigen Punkt $\mathbf{p}$ auf $\mathcal{P}$, d. h. $\mathbf{r}^{\mathrm{T}}\mathbf{p} = -c\mathrm{f}$, gilt die obige Gleichung. Der Abstand zwischen einer verrauschten Messung $\mathbf{y'}_e$ und der Ebene $\mathcal{P}$ ist durch

$$\operatorname{dist}\left(\mathcal{P}, \mathbf{y'}_e\right) = \left| \underbrace{\frac{\mathbf{r}^T}{\|\mathbf{r}\|}}_{:=\mathbf{r'}^{\mathrm{T}}} \left(\hat{\mathbf{y}}_e^T - \mathbf{p}\right) \right| \tag{A.13}$$

definiert. Hierbei wird angenommen, dass die Messungenauigkeiten in $\mathbf{y'}_e$ unabhängig und homogen verteilt sind. Der Vektor $\mathbf{r'}$ stellt den noch unbekannten Normalenvektor der Ebene dar. Unter der Randbedingung $\|\mathbf{r'}\| = 1$ minimiert das Gütekriterium

$$C\left(\mathbf{r'}, \mathbf{p}\right) = \sum_{e \in \mathcal{M}} \left(\mathbf{r'}^{\mathrm{T}} \left(\hat{\mathbf{y}'}_e - \mathbf{p}\right)\right)^2 \rightarrow \min \tag{A.14}$$

dann die Quadratesumme der orthogonalen Abstände aller Messungen zur Ebene. $\mathcal{M}$ drückt hier die Menge aller beobachteten Punkte aus. Mit der Einführung des Zentroiden aller idealen Messungen

$$\overline{\mathbf{y}} = \frac{1}{|\mathcal{M}|} \sum_{e \in \mathcal{M}} \hat{\mathbf{y}'}_e \tag{A.15}$$

kann diese Gleichung umformuliert werden zu

$$\begin{aligned} C\left(\mathbf{r'}, \mathbf{p}\right) &= \sum_{e \in \mathcal{M}} \left(\mathbf{r'}^{\mathrm{T}} \left(\hat{\mathbf{y}'}_e - \overline{\mathbf{y}}\right) - \mathbf{r}^{\mathrm{T}} \left(\mathbf{p} - \overline{\mathbf{y}}\right)\right)^2 \\ &= \sum_{e \in \mathcal{M}} \left(\mathbf{r'}^{\mathrm{T}} \left(\hat{\mathbf{y}'}_e - \overline{\mathbf{y}}\right)\right)^2 + \sum_{e \in \mathcal{M}} \left(\mathbf{r'}^{\mathrm{T}} \left(\mathbf{p} - \overline{\mathbf{y}}\right)\right)^2 + \\ &\quad \underbrace{2\mathbf{r'}^{\mathrm{T}} \sum_{e \in \mathcal{M}} \underbrace{\left(\hat{\mathbf{y}'}_e - \overline{\mathbf{y}}\right)}_{=0} \mathbf{r'}^{\mathrm{T}} \left(\mathbf{p} - \overline{\mathbf{y}}\right)}_{=0} \rightarrow \min. \end{aligned} \tag{A.16}$$

Für jedes $\mathbf{r'}$ ist der erste Summand in obiger Gleichung eine Konstante und der zweite Summand wird minimiert mit $\mathbf{p} = \overline{\mathbf{y}}_e$. Die optimale Ebene $\mathcal{P}$ geht somit durch den Zentroiden $\overline{\mathbf{y}}$. Als Konsequenz kann der erste Teil der Gleichung als gewöhnliches Least-Squares Problem

$$C\left(\mathbf{r'}, \hat{\mathbf{y}'}_e\right) = \sum_{e \in \mathcal{M}} \left(\mathbf{r'}^{\mathrm{T}} \left(\hat{\mathbf{y}'}_e - \overline{\mathbf{y}}\right)\right)^2 = \mathbf{r'}^{\mathrm{T}} \sum_{e \in \mathcal{M}} \left(\hat{\mathbf{y}'}_e \hat{\mathbf{y}'}_e^T\right) \mathbf{r'} \rightarrow \min \tag{A.17}$$

behandelt werden. Die Lösung des Eigenwertproblems liefert den Eigenvektor $\mathbf{r}'$ sowie die zugehörigen Eigenwerte, die (A.17) minimieren. Hieraus kann

$$\mathbf{r} = -\sigma_\Delta^2 \frac{\widehat{\mathbf{r}'}}{\widehat{\mathbf{r}'}_3} \tag{A.18}$$

berechnet werden, wobei $\widehat{\mathbf{r}}$ den Eigenvektor des Eigenwertproblems bezeichnet. Mit (A.12) können danach die Ebenenparameter wie folgt bestimmt werden:

$$a = \mathbf{r}_1 \sqrt{\frac{1-c^2}{\mathbf{r}_1^2+\mathbf{r}_2^2}}\,, \qquad b = \mathbf{r}_2 \sqrt{\frac{1-b^2}{\mathbf{r}_1^2+\mathbf{r}_2^2}}\,, \qquad c = -\frac{1}{\mathrm{f}} \mathbf{r}^{\mathrm{T}} \overline{\mathbf{y}}\,, \tag{A.19}$$

mit der Normierungskonstante

$$\varphi = \frac{\mathrm{b}\, a\, \sigma_x^2}{\mathbf{r}_1}\,. \tag{A.20}$$

## A.3 Schätzen mit unvollständigen Daten

Gegeben sei der Messvektor $\mathbf{Y}$, der im Weiteren die unvollständigen Daten darstellt und durch einen stochastischen Prozess erzeugt wird. Ist es möglich, die Wahrscheinlichkeitsverteilung von $\mathbf{Y}$ durch einen gewünschten Satz von Modellparametern $\boldsymbol{\theta}$ zu charakterisieren, können diese Parameter aus den Daten geschätzt und für eine Anwendung nutzbar gemacht werden. Ein vielverwendetes Verfahren in diesem Zusammenhang ist die Maximum-Likelihood-(ML-)Schätzung. Hierfür wird die sogenannte *Likelihood-Funktion* $P(\mathbf{Y}|\boldsymbol{\theta})$ definiert, die für jede Parameterkonfiguration angibt, wie wahrscheinlich es unter der entsprechenden Verteilung ist, die beobachtete Stichprobe zu erhalten. Die ML-Schätzung bestimmt nun

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \{P(\mathbf{Y}|\boldsymbol{\theta})\} \tag{A.21}$$

als den Parametersatz mit höchster Wahrscheinlichkeit bzgl. der Daten. Für die Ermittlung eines Maximums wird hierbei oftmals zu einer numerisch besser handhabbaren Maximierung der logarithmierten Likelihoodfunktion übergegangen. Man spricht dann von der sogenannten *Log-Likelihood*.

Dieses Schätzverfahren stößt jedoch an seine Grenzen, falls die Verteilung durch mehrere Prozesse erzeugt wird. Die Schwierigkeit besteht darin, diese Prozesse zu identifizieren und jede der Beobachtungen dem sie erzeugenden Prozess zuzuweisen. Hierfür wird die bestehende Beschreibung um den Zufallsprozess $\mathbf{l}$ mit den

Realisationen $\mathbf{l} = (l_1, \ldots, l_N)$ erweitert. Diese Größe stellt die verdeckten, zusätzlich zu bestimmenden, Daten dar. Der komplette Datensatz lautet somit $\mathbb{Y} = \{\mathbf{Y}, \mathbf{l}\}$ mit der Likelihood der vollständigen Daten

$$P(\mathbb{Y}|\boldsymbol{\theta}) = P(\mathbf{Y}, \mathbf{l}|\boldsymbol{\theta}) = P(\mathbf{l}|\mathbf{Y}, \boldsymbol{\theta})P(\mathbf{Y}|\boldsymbol{\theta}). \tag{A.22}$$

Im Unterschied zu der Beschreibung in (A.21) stellen die einzelnen Komponenten der Likelihoodfunktion nun keine skalaren Werte mehr dar, sondern eine Zufallsgröße, die durch $\mathbf{l}$ bedingt ist. Die Log-Likelihood ergibt sich demnach zu

$$\log P(\mathbf{Y}|\boldsymbol{\theta}) = \log P(\mathbb{Y}|\boldsymbol{\theta}) \sum_{\mathrm{l} \in \mathbf{l}} f(\mathrm{l}) - \log P(\mathbf{l}|\mathbf{Y}, \boldsymbol{\theta}) \sum_{\mathrm{l} \in \mathbf{l}} f(\mathrm{l}), \tag{A.23}$$

wobei die einzelnen Terme um einen Ausdruck $f(\mathrm{l})$, welcher die Bedingung[1] $\sum_{\mathrm{l} \in \mathbf{l}} f(\mathrm{l}) = 1$ erfüllt, erweitert wurden. Ersetzt man nun $f(\mathrm{l})$ durch die bedingte Verteilung von l, ergeben sich die jeweiligen Erwartungswerte der Ausdrücke zu

$$\begin{aligned} \log P(\mathbf{Y}|\hat{\boldsymbol{\theta}}) &= \sum_{\mathrm{l} \in \mathbf{l}} \log P(\mathbb{Y}|\hat{\boldsymbol{\theta}}) P(\mathrm{l}|\mathbf{Y}, \boldsymbol{\theta}) - \sum_{\mathrm{l} \in \mathbf{l}} \log P(\mathbf{l}|\mathbf{Y}, \hat{\boldsymbol{\theta}}) P(\mathrm{l}|\mathbf{Y}, \boldsymbol{\theta}) \\ &= \underbrace{\mathrm{E}\left[\log P(\mathbb{Y}|\boldsymbol{\theta})|\mathbf{Y}, \hat{\boldsymbol{\theta}}\right]}_{\mathrm{Q}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})} - \underbrace{\mathrm{E}\left[\log P(\mathbf{l}|\mathbf{Y}, \boldsymbol{\theta})|\mathbf{Y}, \hat{\boldsymbol{\theta}}\right]}_{\mathrm{H}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})}. \end{aligned} \tag{A.24}$$

Der Term $\mathrm{Q}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$ wird als *Kullback-Leibler-Abstand* bzw. Kullback-Leibler-Statistik bezeichnet und kann als Maß für die Ähnlichkeit zweier beliebiger Verteilungen aufgefasst werden. Der zweite Term $\mathrm{H}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$ wird als bedingte Entropie oder auch Äquivokation bezeichnet und soll hier nicht weiter betrachtet werden. Mit obiger Umformung ist es nun möglich, bei gegebenem Parametersatz $\hat{\boldsymbol{\theta}}$ eine neue, aktuelle Schätzung von $\boldsymbol{\theta}$ zu generieren. Es kann gezeigt werden, dass der Ausdruck monoton gegen ein lokales Maximum konvergiert. Für einen Beweis wird auf [Dempster u. a., 1977] verwiesen.

Zur Maximierung der Likelihood in Gleichung (A.24) wertet das EM-Verfahren (engl. *EM=**E**xpecation-**M**aximization*) iterativ nur den ersten Term aus. Hierbei wechselt das Verfahren zu jedem Iterationsschritt $k = (1, \ldots, K)$ zwischen der Bestimmung des Erwartungswertes der vollständigen Daten

$$\mathrm{Q}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^k) = \mathrm{E}\left[\log P(\mathbb{Y}|\boldsymbol{\theta})|\mathbf{Y}, \hat{\boldsymbol{\theta}}^k\right] = \sum_{\mathrm{l} \in \mathbf{l}} \log P(\mathbf{Y}|\boldsymbol{\theta}, \mathrm{l}) P(\mathrm{l}|\mathbf{Y}, \hat{\boldsymbol{\theta}}^k) \tag{A.25}$$

[1] Im kontinuierlichen Fall geht die Summe in ein Integral über und es gilt $\int f(\mathrm{l}) d\mathrm{l} = 1$.

im **E-Schritt** und seiner Maximierung

$$\hat{\boldsymbol{\theta}}^{k+1} = \arg\max_{\boldsymbol{\theta}} \left\{ \mathrm{Q}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{k}) \right\} \tag{A.26}$$

bzgl. $\boldsymbol{\theta}$ im **M-Schritt**, bis keine sichtliche Verbesserung von $\mathrm{Q}(\boldsymbol{\theta}^{k+1}|\hat{\boldsymbol{\theta}}^{k})$ mehr erkennbar ist.

## A.4 Dynamische Zustandsschätzung

Grundlage der dynamischen Zustandsschätzung bildet eine statistische Systembetrachtung, bei der sowohl die *zeitliche Veränderung* des Systemzustandes als auch die zur Verfügung stehenden *Beobachtungsgrößen* mit Unsicherheiten behaftet sind und bei der Schätzung berücksichtigt werden. In diesem Zusammenhang spielt das im Folgenden vorgestellte Bayes-Filter zur rekursiven Zustandsschätzung eine zentrale Rolle. Eine in der Praxis sehr viel verwendete Spezialisierung dieser Formulierung ist das Kalman-Filter, welche im Anschluss daran näher erläutert wird.

### A.4.1 Das Bayes-Filter

In dieser Arbeit sind speziell rekursive Verfahren zur Integration der Systemdynamik in den Schätzprozess von Interesse. Wie in [Stiller u. a., 2009; Krebs, 1999] gezeigt wird, ist mit dem Bayes-Filter hierbei eine der allgemeingültigsten Formen eines rekursiven Schätzers gegeben. Gesucht ist die a-posteriori Verteilungsdichte $P(\boldsymbol{\theta}_t|\mathbf{Y}_{0:t})$ der Zufallsgröße $\boldsymbol{\theta}_t$, basierend auf der fehlerbehafteten Messfolge $\mathbf{Y}_{0:t} = (\mathbf{Y}_0, \ldots, \mathbf{Y}_t)$. Unter Ausnutzung des Satzes von Bayes ergibt sich hierfür die sog. Filterdichte zu

$$P(\boldsymbol{\theta}_t|\mathbf{Y}_{0:t}) = cP(\mathbf{Y}_t|\boldsymbol{\theta}_t, \mathbf{Y}_{0:t-1})P(\boldsymbol{\theta}_t|\mathbf{Y}_{0:t-1}), \tag{A.27}$$

wobei die Normalisierungskonstante $c = P(\mathbf{Y}_t|\mathbf{Y}_{0:t-1})^{-1}$ garantiert, dass $\int P(\boldsymbol{\theta}_t|\mathbf{Y}_{0:t})d\boldsymbol{\theta}_t = 1$ eine Wahrscheinlichkeitsverteilung darstellt. Unter der Annahme zeitlicher Unabhängigkeit der Beobachtungen bei gegebenem Zustand $\boldsymbol{\theta}_t$, vereinfacht sich der erste Faktor zu $P(\mathbf{Y}_t|\boldsymbol{\theta}_t, \mathbf{Y}_{0:t-1}) = P(\mathbf{Y}_t|\boldsymbol{\theta}_t)$. Der zweite Faktor beschreibt die Veränderung der Verteilungsdichte infolge der Streckendynamik und wird deshalb auch als Prädiktions-Verteilungsdichte bezeichnet. Unter Ausnutzung der Kettenregel und des Gesetzes der totalen Wahrscheinlichkeit kann

man diese Größe formal umschreiben:

$$\begin{aligned} P(\boldsymbol{\theta}_t|\mathbf{Y}_{0:t-1}) &= \int P(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\mathbf{Y}_{0:t-1}) d\boldsymbol{\theta}_{t-1} \\ &= \int P(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathbf{Y}_{0:t-1}) P(\boldsymbol{\theta}_{t-1}|\mathbf{Y}_{0:t-1}) d\boldsymbol{\theta}_{t-1} \,. \end{aligned} \tag{A.28}$$

Berücksichtigt man, dass der Zustand einem Markov-Prozess entspringt, lässt sich (A.27) in die endgültige rekursive Gleichung des Bayes-Filters überführen:

$$P(\boldsymbol{\theta}_t|\mathbf{Y}_{0:t}) = cP(\mathbf{Y}_t|\boldsymbol{\theta}_t) \int \underbrace{P(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})}_{f(\boldsymbol{\theta}_t)} \underbrace{P(\boldsymbol{\theta}_{t-1}|\mathbf{Y}_{0:t-1})}_{\hat{\boldsymbol{\theta}}_{t-1}} d\boldsymbol{\theta}_{t-1} \,. \tag{A.29}$$

Es müssen somit nur noch die jeweiligen Modelle zur Beschreibung der Beobachtungen und der Systemdynamik $f(\boldsymbol{\theta}_t)$ unter Verwendung der a-priori Systemkenntnisse spezifiziert werden.

Die eigentliche Rekursion geschieht durch die Integration der Schätzung $\hat{\boldsymbol{\theta}}_{t-1}$ des vorherigen Zeitschritts in folgender Weise: nach Initialisierung des Filters mit $p(\boldsymbol{\theta}_0)$ wird in einem ersten Schritt, dem sog. Prädiktionsschritt, die Schätzung des Zustandes zum Zeitpunkt $t-1$ mit Hilfe von $f(\boldsymbol{\theta}_t)$ nach $t$ projiziert. Dazu wird der Ausdruck hinter dem Integral in Gleichung (A.29), welcher auch als *Chapman-Kolomogoroff-Gleichung* [Chapman, 1928; Kolmogoroff, 1933] bekannt ist, ausgewertet. Im nachfolgenden Innovationsschritt wird die Prädiktion dann durch die aktuelle Beobachtung verbessert.

Wird das Filter mit der wahren Wahrscheinlichkeitsverteilung $p(\boldsymbol{\theta}_0)$ des Zustands initialisiert, so schätzt das Bayes-Filter den Zustand des Systems unter der Markov-Annahme für alle folgenden Zeitschritte korrekt. In der Praxis findet das Bayes-Filter vielfach Anwendung, wobei die einzelnen Ansätze sich grundlegend nur in der Repräsentation der Wahrscheinlichkeitsdichtefunktion $P(\boldsymbol{\theta}_t|\mathbf{Y}_{0:t})$ unterscheiden. Für den Fall eines linearen Systems mit normalverteilten Anfangswerten ist (A.29) in geschlossener Form lösbar und unter der Bezeichnung Kalman-Filter gemeinhin bekannt.

### A.4.2 Das Kalman-Filter

Das Kalman-Filter (KF) ist ein rekursiver Algorithmus zur optimalen Schätzung der Parameter eines dynamischen Systems aus einer Folge von fehlerbehafteten

Beobachtungen. Hierbei wird das betrachtete System als linear und die stochastischen Größen als unimodal und gaußverteilt angenommen, d. h. sie sind vollständig durch die ersten beiden Momente ihre Verteilung (Erwartungswert und Kovarianz) bestimmt.[2]

Für zeitdiskrete Systeme wird der Zustand $\boldsymbol{\theta}_t$ im Allgemeinen vektorwertig angegeben und gehorcht dem linearen (stochastischen) Systemmodell

$$\boldsymbol{\theta}_t = \mathbf{A}_{t-1}\boldsymbol{\theta}_{t-1} + \mathbf{B}_{t-1}\mathbf{u}_{t-1} + \mathbf{q}_{t-1}\,, \tag{A.30}$$

mit Transitionsmatrix $\mathbf{A}_{t-1}$, Eingangsmatrix $\mathbf{B}_{t-1}$, bekanntem Eingangsvektor $\mathbf{u}_{t-1}$ und System- bzw. Prozessrauschen $\mathbf{q}_{t-1}$. Der Zusammenhang zwischen Beobachtung $\mathbf{y}_t$ und Systemmodell wird durch das Beobachtungsmodell der Form

$$\mathbf{y}_t = \mathbf{C}_t\boldsymbol{\theta}_t + \mathbf{e}_t \tag{A.31}$$

gegeben. $\mathbf{C}_t$ bezeichnet hier die Beobachtungsmatrix und $\mathbf{e}_t$ repräsentiert das Beobachtungsrauschen. Die Rauschterme des System- und Beobachtungsmodells werden beim klassischen KF durch mittelwertfreies, normalverteiltes, weißes Rauschen[3] mit Kovarianz $\Sigma_{\mathbf{qq}}$ und $\Sigma_{\mathbf{ee}}$ modelliert, d. h. $\mathbf{q}_t \sim \mathcal{N}(0, \Sigma_{\mathbf{qq}})$ und $\mathbf{e}_t \sim \mathcal{N}(0, \Sigma_{\mathbf{ee}})$.

Da aufgrund der beschriebenen Eigenschaften die Normalverteilungsannahme über die Zeit hinweg erhalten bleibt, beschränkt sich das Filterproblem auf die rekursive Schätzung der ersten beiden Momente der Verteilung, was beim KF prinzipiell in drei Phasen erfolgt:

- ◇ **Initialisierung** des Filters mit $\mathcal{N}(\boldsymbol{\theta}_0, \Sigma_{\mathbf{qq},0})$
- ◇ Propagation des Systemzustands über die Zeit. Dieser Schritt wird als **Prädiktionsschritt** bezeichnet. Die sich hieraus ergebende Verteilung des Zustands wird als a-priori-Schätzung $\mathcal{N}(\hat{\boldsymbol{\theta}}_{t|t-1}, \mathbf{P}_{t|t-1})$ bzgl. der Verteilung $\mathcal{N}(\hat{\boldsymbol{\theta}}_{t-1}, \mathbf{P}_{t-1})$ zum letzten Zeitpunkt bezeichnet. Der Systemzustand $\hat{\boldsymbol{\theta}}_{t|t-1} = \mathrm{E}[\boldsymbol{\theta}_t]$ für den folgenden Zeitschritt ergibt mit Hilfe von Gleichung (A.30)

$$\hat{\boldsymbol{\theta}}_{t|t-1} = \mathbf{A}_{t-1}\hat{\boldsymbol{\theta}}_{t-1} + \mathbf{B}_{t-1}\mathbf{u}_{t-1}\,. \tag{A.32}$$

[2] Sind alle diese Voraussetzungen erfüllt, so liefert das KF optimale Schätzergebnisse, wobei das KF selbst für nicht gaussverteilte Wahrscheinlichkeitsdichten noch immer das beste lineare erwartungstreue Filter minimaler Varianz darstellt.

[3] $\mathrm{E}\left[\mathbf{q}_u\mathbf{q}_v^{\mathrm{T}}\right] = \mathbf{0}$, $\mathrm{E}\left[\mathbf{e}_u\mathbf{e}_v^{\mathrm{T}}\right] = \mathbf{0}$, $\mathrm{E}\left[\mathbf{q}_u\mathbf{e}_v^{\mathrm{T}}\right] = \mathbf{0}$, $u, v \in \mathbb{N}$, $u \neq v$

Die durch die Prädiktion ansteigende Unsicherheit der Schätzung wird durch Anpassung der Kovarianzmatrix des Schätzfehlers $(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}_t)$ erreicht. Unter Verwendung des Systemmodells aus Gleichung (A.30) folgt hieraus

$$\begin{aligned} \mathbf{P}_{t|t-1} &= \mathrm{E}\left[(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}_t)(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}_t)^{\mathrm{T}}\right] \\ &= \mathbf{A}_{t-1}\mathrm{E}\left[(\hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}_t)(\hat{\boldsymbol{\theta}}_{t-1} - \boldsymbol{\theta}_t)^{\mathrm{T}}\right]\mathbf{A}_{t-1}^{\mathrm{T}} + \Sigma_{\mathbf{ee}} \\ &= \mathbf{A}_{t-1}\mathbf{P}_{t-1}\mathbf{A}_{t-1}^{\mathrm{T}} + \Sigma_{\mathbf{ee}} \end{aligned} \tag{A.33}$$

Nach der Prädiktion erfolgt der Iterationsschritt mit $t := t+1$.

◇ Im **Innovationsschritt** wird der a-priori-Schätzwert $\hat{\boldsymbol{\theta}}_{t|t-1}$ des Systemzustands mit Hilfe der aktuellen Beobachtung $\mathbf{y}_t$ korrigiert. Der sich hieraus ergebende a-posteriori-Schätzwert $\hat{\boldsymbol{\theta}}_t$ der Systemzustände ergibt sich zu

$$\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{K}_t\left[\mathbf{y}_t - \mathbf{C}_t\hat{\boldsymbol{\theta}}_{t|t-1}\right], \tag{A.34}$$

wobei für den Augenblick angenommen werden soll, die Größe $\mathbf{K}_t$ sei bekannt. Der nach der Innovation vorhandene Schätzfehler $(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)$ führt dann unter Ausnutzung von Gleichung (A.34) zu folgendem Ausdruck für die aktualisierte Kovarianzmatrix

$$\begin{aligned} \mathbf{P}_t &= \mathrm{E}\left[(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t)^{\mathrm{T}}\right] \\ &= [\mathrm{I} - \mathbf{K}_t\mathbf{C}_t]\,\mathbf{P}_{t|t-1}\,[\mathrm{I} - \mathbf{K}_t\mathbf{C}_t]^{\mathrm{T}} + \mathbf{K}_t\Sigma_{\mathbf{ee}}\mathbf{K}_t^{\mathrm{T}}. \end{aligned} \tag{A.35}$$

Es sei kurz in Erinnerung gerufen, dass das KF einen Minimum-Varianz Schätzer darstellt, also gerade die Summe der Varianzen der geschätzten Parameter minimiert. Als Kostenfunktion kann somit die Spur der Kovarianzmatrix $\mathbf{P}_t$ verwendet werden. Zur Bestimmung der unbekannten Matrix $\mathbf{K}_t$ wird genau diese Kostenfunktion nach der gesuchten Größe abgeleitet und zu null gesetzt:

$$\frac{d\mathrm{Spur}(\mathbf{P}_t)}{d\mathbf{K}_t} = -2\left[\mathbf{C}_t\mathbf{P}_{t|t-1}\right]^{\mathrm{T}} + 2\mathbf{K}_t\left[\mathbf{C}_t\mathbf{P}_{t|t-1}\mathbf{C}_t^{\mathrm{T}} + \Sigma_{\mathbf{ee}}\right] = \mathbf{0}. \tag{A.36}$$

Die *Kalman-Verstärkungsmatrix* (engl. *Kalman gain*) ergibt sich dann durch Auflösen der obigen Gleichung nach $\mathbf{K}_t$ zu

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{C}_t^{\mathrm{T}}\left[\mathbf{C}\mathbf{P}_{t|t-1}\mathbf{C}_t^{\mathrm{T}} + \Sigma_{\mathbf{ee}}\right]^{-1}. \tag{A.37}$$

Durch Einsetzen der Kalman-Verstärkungsmatrix in Gleichung (A.35) erhält man die mathematisch äquivalente, allerdings numerisch weniger robustere Form der *Kovarianzmatrix-Korrektur*

$$\mathbf{P}_t = [\mathrm{I} - \mathbf{K}_t \mathbf{C}_t]\, \mathbf{P}_{t|t-1}\,. \tag{A.38}$$

Während die Initialisierung nur einmal durchlaufen wird, wechseln die genannten Schritte Prädiktion und Innovation im gesamten Prozess miteinander ab. Man spricht daher auch von einem Prädiktor-Korrektor-Verfahren.

Mit zunehmender Länge der Messreihe nähert sich die Schätzung $\mathcal{N}(\hat{\boldsymbol{\theta}}_t, \mathbf{P}_t)$ der tatsächlichen Verteilung beliebig genau an, d. h. es handelt sich um einen erwartungstreuen und konsistenten Schätzer mit minimaler Varianz. Aufgrund dieser Schätzeigenschaften ist das KF ein optimales lineares Filter. Selbst verallgemeinerte nichtlineare Filter liefern für das hier betrachtete lineare Zustandsraummodell mit normalverteilten Variablen keine besseren Ergebnisse. Im Gegensatz zu anderen (rekursiven) linearen Schätzern, erlaubt das KF auch die Behandlung von Problemen mit korrelierten Rauschkomponenten, wie sie in der Praxis häufig anzutreffen sind.

### A.4.3 Das erweiterte Kalman-Filter

Für lineare stochastische Systeme stellt das oben skizzierte Verfahren einen optimale Schätzer dar. Jedoch führen in der Realität schon einfache Schätzaufgaben zu nichtlinearen Zustands- oder Beobachtungsgleichungen. Abhilfe schaffen hier u. a. nichtlineare Erweiterungen des Kalman-Filters wie das bereits in den 60-er Jahren entwickelte erweiterte Kalman-Filter (EKF), welches eine Linearisierung um den aktuellen Systemzustand vornimmt. Die Herleitung des EKF, welches auch als direktes oder Total-State-Space-Kalman-Filter bezeichnet wird, folgt dabei den Ausführungen in [Simon 2006]. Die Dynamik des zeitdiskreten Systemmodells wird hier durch die nichtlineare Funktion $\mathbf{f}$ beschrieben:

$$\boldsymbol{\theta}_t = \mathbf{f}(\boldsymbol{\theta}_{t-1}, \mathbf{u}_{t-1}, \mathbf{q}_{t-1})\,. \tag{A.39}$$

Die im Allgemeinen ebenfalls meist nichtlineare Beobachtungsgleichung wird beschrieben durch

$$\mathbf{y}_t = \mathbf{h}(\boldsymbol{\theta}_t, \mathbf{e}_t)\,. \tag{A.40}$$

Die Variablenbezeichnungen aus (A.30)-(A.31) und deren entsprechenden Eigenschaften können direkt aus der Beschreibung des KF übernommen werden. Ebenfalls analog zum KF erfolgt die Zustandsschätzung innerhalb des EKF in alternierender Abfolge der beiden Schritte Prädiktion und Innovation:

◇ Der **Prädiktionsschritt** umfasst, wie beim KF , die Propagation der Systemzustände in den nächsten Zeitschritt. Dazu muss zuerst das nichtlineare Systemmodell durch eine Taylor-Approximation erster Ordnung der Funktion $\mathbf{f}$ um den bekannten Systemzustand $\hat{\boldsymbol{\theta}}_{t-1}$ linearisiert werden, also

$$\boldsymbol{\theta}_t \approx \mathbf{f}(\hat{\boldsymbol{\theta}}_{t-1}, \mathbf{u}_{t-1}, \mathbf{0}) + \mathbf{F}_{\boldsymbol{\theta}} \cdot \left(\boldsymbol{\theta}_{t-1} - \hat{\boldsymbol{\theta}}_{t-1}\right) + \mathbf{F}_{\mathbf{q}} \cdot \mathbf{q}_{t-1}\,, \tag{A.41}$$

wobei $\mathbf{F}_{\boldsymbol{\theta}}$ und $\mathbf{F}_{\mathbf{q}}$ die Jakobimatrizen der Funktion $\mathbf{f}$ nach den Parametern $\boldsymbol{\theta}$ und $\mathbf{q}$ bezeichnen:

$$\mathbf{F}_{\boldsymbol{\theta}} := \left.\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}_{t-1}, \mathbf{u}_{t-1}, \mathbf{0}}, \quad \mathbf{F}_{\mathbf{q}} := \left.\frac{\partial \mathbf{f}}{\partial \mathbf{q}}\right|_{\hat{\boldsymbol{\theta}}_{t-1}, \mathbf{u}_{t-1}, \mathbf{0}}. \tag{A.42}$$

Die allgemeine Systemgleichung (A.41) kann weiter vereinfacht werden zu

$$\boldsymbol{\theta}_t = \mathbf{F}_{\boldsymbol{\theta}}\, \boldsymbol{\theta}_{t-1} + \tilde{\mathbf{u}}_{t-1} + \tilde{\mathbf{q}}_{t-1}\,, \tag{A.43}$$

wobei

$$\tilde{\mathbf{u}}_{t-1} = \mathbf{f}(\hat{\boldsymbol{\theta}}_{t-1}, \mathbf{u}_{t-1}, \mathbf{0}) - \mathbf{F}_{\boldsymbol{\theta}}\, \hat{\boldsymbol{\theta}}_{t-1} \tag{A.44}$$

eine bekannte, virtuelle Stellgröße repräsentiert und

$$\tilde{\mathbf{q}}_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{F}_{\mathbf{q}} \Sigma_{\mathbf{ee}} \mathbf{F}_{\mathbf{q}}^T)\,. \tag{A.45}$$

Die nichtlineare Systemgleichung (A.39) kann somit durch eine lineare Approximation (siehe (A.43)-(A.45)) ersetzt werden, welche der Standardform des KF entspricht. Hierdurch lässt sich der Prädiktionsschritt des EKF analog zum Prädiktionsschritt des KF formulieren:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{t|t-1} &= \mathbf{f}(\hat{\boldsymbol{\theta}}_{t-1}, \mathbf{u}_{t-1}, \mathbf{0}) \\ \mathbf{P}_{t|t-1} &= \mathbf{F}_{\boldsymbol{\theta}} \mathbf{P}_{t-1} \mathbf{F}_{\boldsymbol{\theta}}^T + \mathbf{F}_{\mathbf{q}} \Sigma_{\mathbf{ee}} \mathbf{F}_{\mathbf{q}}^T\,. \end{aligned} \tag{A.46}$$

◇ Im **Innovationsschritt** des EKF muss ebenfalls das nichtlineare Beobachtungsmodell (A.40) durch eine Taylor-Approximation erster Ordnung der Funktion $\mathbf{h}$ um den Systemzustand $\hat{\boldsymbol{\theta}}_{t-1}$ linearisiert werden:

$$\tilde{\mathbf{y}}_t \approx \mathbf{h}(\hat{\boldsymbol{\theta}}_t, \mathbf{0}) + \mathbf{H}_{\boldsymbol{\theta}} \cdot \left(\boldsymbol{\theta}_t - \hat{\boldsymbol{\theta}}_t\right) + \mathbf{H}_{\mathbf{e}} \cdot \mathbf{e}_t\,, \tag{A.47}$$

wobei $\mathbf{H}_{\boldsymbol{\theta}}$ und $\mathbf{H}_{\mathbf{e}}$ wieder die Jakobimatrizen der Funktion $\mathbf{h}$ nach den Parametern $\boldsymbol{\theta}$ und $\mathbf{q}$ bezeichnen:

$$\mathbf{H}_{\boldsymbol{\theta}} := \left.\frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}_{t-1},\mathbf{0}}, \quad \mathbf{H}_{\mathbf{e}} := \left.\frac{\partial \mathbf{h}}{\partial \mathbf{e}}\right|_{\hat{\boldsymbol{\theta}}_{t-1},\mathbf{0}}. \tag{A.48}$$

Hieraus ergibt sich die Beobachtungsgleichung durch weitere Vereinfachungen zu

$$\tilde{\mathbf{y}}_t \approx \mathbf{H}_{\boldsymbol{\theta}} \cdot \boldsymbol{\theta}_t + \mathbf{h}(\hat{\boldsymbol{\theta}}_t, \mathbf{0}) - \mathbf{H}_{\boldsymbol{\theta}} \cdot \hat{\boldsymbol{\theta}}_t + \tilde{\mathbf{e}}_t, \tag{A.49}$$

mit dem Beobachtungsrauschen

$$\tilde{\mathbf{e}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{H}_{\mathbf{e}} \Sigma_{\mathbf{ee}} \mathbf{H}_{\mathbf{e}}^T), \tag{A.50}$$

Abweichend vom linearen KF berechnet sich dann die a-posteriori Verteilung des Systemzustandes aus der gewichteten Differenz des Beobachtungsvektors $\tilde{\mathbf{y}}_t$ und der erwarteten Beobachtung zu

$$\begin{aligned}
\mathbf{K}_t &= \mathbf{P}_{t|t-1} \mathbf{H}_{\boldsymbol{\theta}}^T \left[ \mathbf{H}_{\boldsymbol{\theta}} \mathbf{P}_{t|t-1} \mathbf{H}_{\boldsymbol{\theta}}^T + \mathbf{H}_{\mathbf{q}} \Sigma_{\mathbf{ee}} \mathbf{H}_{\mathbf{q}} \right]^{-1} \\
\hat{\boldsymbol{\theta}}_t &= \hat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{K}_t \left[ \tilde{\mathbf{y}}_t - \mathbf{h}(\hat{\boldsymbol{\theta}}_t, \mathbf{0}) \right] \\
\mathbf{P}_t &= \left[ \mathrm{I} - \mathbf{K}_t \mathbf{H}_{\boldsymbol{\theta}} \right] \mathbf{P}_{t|t-1}.
\end{aligned} \tag{A.51}$$

# Literaturverzeichnis


[Abend u. a. 1965] ABEND, K. ; HARLEY, T.J. ; KANAL, L.N.: Classification of Binary Random Patterns. In: *IEEE Transactions on Information Theory* 11 (1965), October, Nr. 4, S. 538–544

[Adiv 1985] ADIV, G.: Determining Three-Dimensional Motion and Structure from Optical Flow Generated by Several Moving Objects. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 7 (1985), July, Nr. 4, S. 384–401

[Agrawal u. a. 2005] AGRAWAL, M. ; KONOLIGE, K. ; IOCCHI, L.: Real-Time Detection of Independent Motion using Stereo. In: *IEEE Workshop on Motion and Video Computing* Bd. 2, 2005, S. 207–214

[Al-Naffouri 2007] AL-NAFFOURI, T.Y.: An EM-Based Forward-Backward Kalman Filter for the Estimation of Time-Variant Channels in OFDM. In: *IEEE Transactions on Signal Processing* 55 (2007), July, Nr. 7, S. 3924–3930

[Appleton u. Talbot 2006] APPLETON, B. ; TALBOT, H.: Globally Minimal Surfaces by Continuous Maximal Flows. In: *Pattern Analysis and Machine Intelligence* 28 (2006), Nr. 1, S. 106–118

[Aschwanden u. Guggenbuhl 1993] ASCHWANDEN, P. ; GUGGENBUHL, W.: Experimental Results from a Comparative Study on Correlation-Type Registration Algorithms. In: *Robust Computer Vision* (1993), S. 268–289

[Bachmann 2009] BACHMANN, A.: Applying Recursive EM to Scene Segmentation. In: *Deutsche Arbeitsgemeinschaft für Mustererkennung DAGM e.V.*, Springer-Verlag, 2009 (LNCS 5748), S. 512–521

[Bachmann u. Balthasar 2008] BACHMANN, A. ; BALTHASAR, M.: Context-Aware Object Priors. In: *International Conference on Intelligent Robots and Systems, Workshop on Planning, Perception and Navigation for Intelligent Vehicles*. Nice, France, September 2008

[Bachmann u. Dang 2006a] BACHMANN, A. ; DANG, T.: Detektion bewegter Objekte unter Berücksichtigung räumlicher und zeitlicher Konsistenz. In: *Workshop Fahrerassistenzsysteme FAS*, 2006 ( 4), S. 70–77

[Bachmann u. Dang 2006b] BACHMANN, A. ; DANG, T.: Multiple Object Detection under the Constraint of Spatiotemporal Consistency. In: *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*, 2006, S. 295–300

[Bachmann u. Dang 2008] BACHMANN, A. ; DANG, T.: Improving Motion-Based Object Detection by Incorporating Object-Specific Knowledge. In: *International Journal of Intelligent Information and Database Systems* 2 (2008), Nr. 2, S. 258–276

[Bachmann u. Kuehne 2009] BACHMANN, A. ; KUEHNE, H.: An Iterative Scheme for Motion-Based Scene Segmentation. In: *IEEE International Conference on Computer Vision; Workshop on Dynamical Vision*, 2009, S. 735–742

[Bachmann u. Lulcheva 2009a] BACHMANN, A. ; LULCHEVA, I.: Bayesian Scene Segmentation Incorporating Motion Constraints and Category-Specific Information. In: *International Conference on Computer Vision Theory and Applications*, 2009, S. 291–298

[Bachmann u. Lulcheva 2009b] BACHMANN, A. ; LULCHEVA, I.: Combining Low-Level Segmentation with Relational Classification. In: *IEEE International Conference on Computer Vision; IEEE Workshop on Visual Surveillance*, 2009, S. 1216–1221

[Badino u. a. 2006] BADINO, H. ; FRANKE, U. ; RABE, C. ; GEHRIG, S.: Stereovision based detection of moving objects under strong camera motion. In: *International Conference on Computer Vision Theory and Applications*, 2006

[Balthasar 2007] BALTHASAR, M.: *Ein Ansatz zur szenenspezifischen Objektklassifikation unter Verwendung kontextbehafteter Bildmerkmale*, Universität Karlsruhe (TH), Institut für Mess- und Regelungstechnik, 76 131 Karlsruhe, Diplomarbeit, 2007

[Bar-Shalom 1987] BAR-SHALOM, Y.: *Tracking and data association*. San Diego, CA, USA : Academic Press Professional, Inc., 1987. – ISBN 0–120–79760–7

[Barron u. a. 1992] BARRON, J.L. ; FLEET, D.J. ; BEAUCHEMIN, S.S. ; BURKITT, T.A.: Performance of Optical Flow Techniques, 1992 (1), S. 43–77

[Baumela u. a. 2000] BAUMELA, L. ; AGAPIT, L. ; BUSTO, P. ; REID, I.: Motion estimation using the differential epipolar equation. In: *International Conference on Pattern Recognition* Bd. 3, 2000, S. 840–843

[Bay u. a. 2008] BAY, H. ; ESS, A. ; TUYTELAARS, T. ; GOOL, L. V.: Speeded-Up Robust Features (SURF). In: *Computer Vision and Image Understanding* 110 (2008), Nr. 3, S. 346–359

[Bell u. Cathey 1993] BELL, B.M. ; CATHEY, F.W.: The iterated Kalman filter update as a Gauss-Newton method. In: *IEEE Transactions on Automatic Control* 38 (1993), Feb, Nr. 2, S. 294–297

[Bergen u. a. 1992] BERGEN, J.R. ; ANANDAN, P. ; HANNA, K.J. ; HINGORANI, R.: Hierarchical Model-Based Motion Estimation. In: *Proceedings of the European Conference on Computer Vision*, Springer-Verlag, 1992, S. 237–252

[Bertozzi u. a. 1997] BERTOZZI, M. ; BROGGI, A. ; CASTELLUCIO, S.: A real-time oriented system for vehicle detection. In: *Journal of System Architecture* 43 (1997), Nr. 1–5, S. 317–325

[Besag 1974] BESAG, J.: Spatial Interaction and the Statistical Analysis of Lattice Systems. In: *Journal of the Royal Statistical Society* (1974), Nr. 2, S. 192–236

[Besag 1986] BESAG, J.: On the statistical analysis of dirty pictures (with discussion). In: *Journal of the Royal Statistical Society* 48 (1986), S. 259–302

[Bhat u. Nayar 1998] BHAT, D.N. ; NAYAR, S.K.: Ordinal Measures for Image Correspondence. In: *Pattern Analysis and Machine Intelligence* 20 (1998), S. 415–423

[Biederman u. a. 1982] BIEDERMAN, I. ; MEZZANOTTE, R. J. ; RABINOWITZ, J. C.: Scene perception: detecting and judging objects undergoing relational violations. In: *Cogintive Psychology*, 1982 ( 14), S. 143–177

[Birchfield u. Tomasi 1998] BIRCHFIELD, S. ; TOMASI, C.: A pixel dissimilarity measure that is insensitive to image sampling. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20 (1998), S. 401–406

[Birchfield u. Tomasi 1999] BIRCHFIELD, S. ; TOMASI, C.: Multiway cut for stereo and motion with slanted surfaces. In: *Proceedings of the IEEE International Conference on Computer Vision* Bd. 1, 1999, S. 489–495 vol.1

[Bishop 2006] BISHOP, C.M.: *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA : Springer-Verlag New York, Inc., 2006

[Black u. Rangarajan 1996] BLACK, M.J. ; RANGARAJAN, A.: On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. In: *International Journal of Computer Vision* 19 (1996), Nr. 1, S. 57–91

[Blake u. Zisserman 1987] BLAKE, A. ; ZISSERMAN, A.: *Visual reconstruction*. Cambridge, Massachusetts, MIT Press, 1987

[Boykov u. Kolmogorov 2001] BOYKOV, Y. ; KOLMOGOROV, V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. In: *Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2001, S. 359–374

[Boykov u. Kolmogorov 2004] BOYKOV, Y. ; KOLMOGOROV, V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2004), September, Nr. 9, S. 1124–1137

[Boykov u. a. 2001] BOYKOV, Y. ; VEKSLER, O. ; ZABIH, R.: Fast Approximate Energy Minimization via Graph Cuts. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001), November, Nr. 11, S. 1222–1239

[Broggi u. a. 2005] BROGGI, A. ; CARAFFI, C. ; ISABELLA, R.F. ; GRISLERI, P.: Obstacle Detection with Stereo Vision for Off-Road Vehicle Navigation. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Patern Recognition*, 2005, S. 65–73

[Brown u. a. 2003] BROWN, M.Z. ; BURSCHKA, D. ; HAGER, G.D.: Advances in computational stereo. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (2003), Aug., Nr. 8, S. 993–1008

[Brox u. a. 2006] BROX, T. ; BRUHN, A. ; WEICKERT, J.: Variational motion segmentation with level sets. In: LEONARDIS, A. (Hrsg.) ; BISCHOF, H. (Hrsg.) ; PINZ, A. (Hrsg.): *European Conference on Computer Vision* Bd. 3951, Springer, May 2006 (LNCS), S. 471–483

[Bruhn u. a. 2005] BRUHN, A. ; WEICKERT, J. ; KOHLBERGER, T. ; SCHNOERR, C.: Discontinuity-Preserving Computation of Variational Optic Flow in Real-Time. In: *Scale-Space*, 2005, S. 279–290

[Bruss u. Horn 1981] BRUSS, A. ; HORN, B.: Passive navigation. In: *Computer Vision, Graphics and Image Processing* 21 (1981), S. 3–20

[Chang u. a. 1997] CHANG, M.M. ; TEKALP, A.M. ; SEZAN, M.I.: Simultaneous Motion Estimation and Segmentation. In: *IEEE Transactions on Image Processing* 6 (1997), Nr. 9, S. 1326–1333

[Chapman 1928] CHAPMAN, S.: On the Brownian Displacements and Thermal Diffusion of Grains Suspended in a Non-Uniform Fluid. In: *Proceedings of the Royal Society* 119 (1928), S. 34–60

[Chou u. Brown 1990] CHOU, P.B. ; BROWN, C.M.: The theory and practice of Bayesian image labeling. In: *International Journal of Computer Vision* 4 (1990), S. 185–210

[Cormen u. a. 2001] CORMEN, T.H. ; LEISERSON, C.E. ; RIVEST, R.L. ; STEIN, C.: *Introduction to Algorithms*. MIT Press and McGraw-Hill, 2001

[Cremers u. Soatto 2005] CREMERS, D. ; SOATTO, S.: Motion Competition: A variational framework for piecewise parametric motion segmentation. In: *International Journal of Computer Vision* 62 (2005), May, Nr. 3, S. 249–265

[Csurka u. Bouthemy 1999] CSURKA, G. ; BOUTHEMY, P.: Direct identification of moving objects and background from 2D motion models. In: *Proceedings of the IEEE International Conference on Computer Vision* Bd. 1, 1999, S. 566–571 vol.1

[Cumani u. Guiducci 2008] CUMANI, A. ; GUIDUCCI, A.: Fast stereo-based visual odometry for rover navigation. In: *World Scientific and Engineering Academy and Society* 7 (2008), Nr. 7, S. 648–657

[Dahlhaus u. a. 1994] DAHLHAUS, E. ; JOHNSON, D.S. ; PAPADIMITRIOU, C.H. ; SEYMOUR, P.D. ; YANNAKAKIS, M.: The Complexity of Multiterminal Cuts. In: *Society for Industrial and Applied Mathematics* 23 (1994), Nr. 4, S. 864–894

[Dang 2007] DANG, T.: *Kontinuierliche Selbstkalibrierung von Stereokameras*, Universtät Karlsruhe; Fakultät für Maschinenbau; Institut für Mess- und Regelungstechnik mit Maschinenlaboratorium (MRT), Diss., Maerz 2007. – Schriftenreihe. Institut für Mess- und Regelungstechnik, Universität Karlsruhe (TH) ; 8

[Dang u. Hoffmann 2005] DANG, T. ; HOFFMANN, C.: Fast Object Hypotheses Generation Using 3D Position and 3D Motion. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2005

[Dang u. Stiller 2009] DANG, T. ; STILLER, C.: Kontinuierliche Selbstkalibrierung von Stereokamera. In: *tm - Technisches Messen* 76 (2009), Nr. 4, S. 167–174

[Dempster u. a. 1977] DEMPSTER, A.P. ; LAIRD, N.M ; RUBIN, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. In: *Journal of the Royal Statistical Society* B 39 (1977), Nr. 1, S. 1–38

[Derin 1986] *Kapitel* 11-The Use of Gibbs Distributions In Image Processing. In: DERIN, H.: *Communications and Networks*. Springer-Verlag, New York, 1986, S. 266–298

[Eisenbeiss 2007] EISENBEISS, E.: *Bestimmung der Bewegungsparameter von Objekten aus Stereobilddaten mit dem Iterative Closest Point Algorithmus*, Universität Karlsruhe (TH), Institut für Mess- und Regelungstechnik, 76 131 Karlsruhe, Diplomarbeit, 2007

[Elias 2003] ELIAS, R.: Clustering Points In nD Space Through Hierarchical Structures. In: *Canadian Conference on Electrical and Computer Engineering* 3 (2003), S. 2079–2081

[Ess u. a. 2009] ESS, A. ; LEIBE, B. ; SCHINDLER, K. ; GOOL, L. V.: Moving Obstacle Detection in Highly Dynamic Scenes. In: *IEEE International Conference on Robotics and Automation*, 2009, S. 4451–4458

[Fang u. Huang 1984] FANG, J.Q. ; HUANG, T.S.: Solving three-dimensional small-rotation motion equations: Uniqueness, algorithms, and numerical results. In: *Computer Vision, Graphics, and Image Processing* 26 (1984), Nr. 2, S. 183–206

[Felzenszwalb u. Huttenlocher 2006] FELZENSZWALB, P. F. ; HUTTENLOCHER, D.P.: Efficient Belief Propagation for Early Vision. In: *International Journal of Computer Vision* 70 (2006), Nr. 1, S. 41–54

[Feng u. Perona 1998] FENG, X. ; PERONA, P.: Scene Segmentation from 3D Motion. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA : IEEE Computer Society, 1998, S. 225–231

[Ferrari u. a. 1995] FERRARI, P. ; FRIGESSI, A. ; GONZAGA, P.: Fast Approximate MAP Restoration of Multicolor Images. In: *Journal of the Royal Statistical Society* 57 (1995), Nr. 3, S. 485–500

[Fischler u. Bolles 1981] FISCHLER, M.A. ; BOLLES, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. In: *Communications of the ACM* 24 (1981), S. 381–395

[Fischler u. Firschein 1987] FISCHLER, M.A (Hrsg.) ; FIRSCHEIN, O. (Hrsg.): *Readings in computer vision: issues, problems, principles, and paradigms*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1987. – ISBN 0–934613–33–8

[Ford u. Fulkerson 1956] FORD, L.R. ; FULKERSON, D.R.: Maximal flow through a network. In: *Canadian Journal of Mathematics* 8 (1956), S. 399–404

[Franke u. Heinrich 2002] FRANKE, U. ; HEINRICH, S.: Fast obstacle detection for urban traffic situations. In: *IEEE Transactions on Intelligent Transportation Systems* 3 (2002), Sep, Nr. 3, S. 173–181

[Franke u. a. 2005] FRANKE, U. ; RABE, C. ; BADINO, H. ; GEHRIG, S.: 6D-Vision: Fusion of Stereo and Motion for Robust Environment Perception. In: *Deutsche Arbeitsgemeinschaft für Mustererkennung DAGM e.V.*, 2005, S. 216–223

[Fua 1993] FUA, P.: A Parallel Stereo Algorithm that Produces Dense Depth Maps and Preserves Image Features. In: *Machine Vision and Applications* 6 (1993), Nr. 1, S. 35–49. – CVLAB-ARTICLE-1993-001

[Gauvrit u. a. 1997] GAUVRIT, H. ; CADRE, J.P. L. ; JAUFFRET, C.: A formulation of multitarget tracking as an incomplete data problem. In: *IEEE Transactions on Aerospace and Electronic Systems* 33 (1997), Oct., Nr. 4, S. 1242–1257

[Geman u. Geman 1984] GEMAN, S. ; GEMAN, D.: Stochstic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. In: *IEEE Transaction on Pattern Analysis and Machine Intelligence* Bd. 6, 1984, S. 721–741

[Gibson 1950] GIBSON, J.J.: *The Perception of the Visual World*. Greenwood Pub Group; Boston: Houghton Mifflin, 1950. – 235 S.

[Goldberg u. Tarjan 1990] GOLDBERG, A. V. ; TARJAN, R.E.: Finding Minimum-Cost Circulations by Successive Approximation. In: *Mathematics of Operations Research* 15 (1990), Nr. 3, S. 430–466

[Green 1990] GREEN, P.: On use of the EM algorithm for penalized likelihood estimation. In: *Journal of the Royal Statistical Society* 52 (1990), Nr. 3, S. 443–452

[Griffeath 1976] *Kapitel* Introduction to random field. In: GRIFFEATH, D.: *Denumerable Markov Chains*. 2. Springer-Verlag, New York, 1976, S. 721–741

[Grimmett 1973] GRIMMETT, R.: A Theorem about Random Fields. In: *Bulletin of the London Mathematical Society* 5 (1973), Nr. 3, S. 81–84

[Hammersley u. Clifford 1971] HAMMERSLEY, J.M. ; CLIFFORD, P.: *Markov field on finite graphs and lattices*. Berkeley preprint, 1971

[Hanna 1991] HANNA, K.J.: Direct multi-resolution estimation of ego motion and structure from motion. In: *Proceedings of the IEEE Workshop on Visual Motion, 1991*, 1991, S. 156–162

[Harris u. Stephens 1988] HARRIS, C. ; STEPHENS, M.J.: A combined corner and edge detector. In: *Alvey Vision Conference*. Manchester, 1988, S. 147–151

[Hartley u. Zisserman 2004] HARTLEY, R.I. ; ZISSERMAN, A.: *Multiple View Geometry in Computer Vision*. 2. Cambridge University Press, 2004

[Haußecker u. Spies 1999] *Kapitel* Motion. In: HAUSSECKER, H. ; SPIES, H.: *Handbook of Computer Vision and Applications*. 1999, S. 309–396

[Heeger u. Jepson 1992] HEEGER, D.J. ; JEPSON, A.D.: Subspace methods for recovering rigid motion I: algorithm and implementation. In: *International Journal of Computer Vision* 7 (1992), Nr. 2, S. 95–117

[Hoffmann 2007] HOFFMANN, C.: *Fahrzeugdetektion durch Fusion monoskopischer Videomerkmale*, Institut für Mess- und Regelungstechnik mit Maschinenlaboratorium (MRT), Universitätsverlag Karlsruhe, Karlsruhe, Schriftenreihe. Institut für Mess- und Regelungstechnik, Universität Karlsruhe (TH) ; 7, 2007

[Horn 1987] HORN, B.: Closed form solution of absulute orientation using unit quaternions. In: *Journal of the Optical Society of America* 4 (1987), S. 629–642

[Horn u. Schunck 1981] HORN, B.K.P. ; SCHUNCK, B.G.: Determining optical flow. In: *Artificial Intelligence* 17 (1981), S. 185–203

[Horn u. a. 2006] HORN, J. ; BACHMANN, A. ; DANG, T.: A Fusion Approach for Image-Based Measurement of Speed Over Ground. In: *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006, S. 261–266

[Horn u. a. 2007] HORN, J. ; BACHMANN, A. ; DANG, T.: Stereo Vision Based Ego Motion Estimation with Sensor Supported Subset Validation. In: *IEEE Intelligent Vehicles Symposium*, 2007, S. 741–748

[Howard 2008] HOWARD, A.: Real-time stereo visual odometry for autonomous ground vehicles. In: *IEEE Intelligent Robots and Systems*, 2008, S. 3946–3952

[Huber 2004] HUBER, P.J.: *Robust Statistics*. John Wiley & Sons, 2004

[Huguet u. Devernay 2007] HUGUET, F. ; DEVERNAY, F.: A Variational Method for Scene Flow Estimation from Stereo Sequences. In: *International Conference on Computer Vision*, 2007, S. 1–7

[Irani u. Anandan 2000] IRANI, M. ; ANANDAN, P.: About Direct Methods. In: *IEEE International Conference on Computer Vision*, Springer-Verlag, 2000, S. 267–277

[Ishikawa 2003] ISHIKAWA, H.: Exact optimization for Markov random fields with convex priors. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25 (2003), Oct., Nr. 10, S. 1333–1336

[Ishikawa u. Geiger 1999] ISHIKAWA, H. ; GEIGER, D.: Mapping image restoration to a graph problem. In: *Nonlinear Signal and Image Processing*, 1999, S. 189–193

[Ising 1925] ISING, E.: Beitrag zur Theorie des Ferromagnetismus. In: *Zeitschrift für Physik* 31 (1925), Nr. 1, S. 253–258

[Jazwinski 1970] JAZWINSKI, Andrew H.: *Stochastic Processes and Filtering Theory*. New York : Academic Press, 1970

[Jähne 2005] JÄHNE, B.: *Digitale Bildverarbeitung*. Bd. 6. Springer-Verlag, 2005

[Kalinke u. Tzokamkas 1998] KALINKE, T. ; TZOKAMKAS, C.: A texture-based object detection and an adaptive model-based classification. In: *IEEE Intelligent Vehicles Symposium*, 1998, S. 143–148

[Kanal 1980] KANAL, L.N.: Markov mesh models. In: *Computer Graphics and Image Processing* 12 (1980), Nr. 4, S. 371–375

[Kato u. a. 1995] KATO, Z. ; ZERUBIA, J. ; BERTHOD, M.: Unsupervised parallel image classification using a hierarchical Markovian model. In: *IEEE International Conference on Computer Vision* 0 (1995), S. 169

[Kitt 2008] KITT, B.: *Ein Ansatz zur quantitativen und qualitativen Bestimmung der Bewegung mehrerer Objekte in Bildsequenzen*, Universität Karlsruhe (TH), Institut für Mess- und Regelungstechnik, 76 131 Karlsruhe, Diplomarbeit, 2008

[Klappstein u. a. 2008] KLAPPSTEIN, J. ; VAUDREY, T. ; RABE, C. ; WEDEL, A. ; KLETTE, R.: Moving Object Segmentation Using Optical Flow and Depth Information. In: *Proceedings of the Pacific Rim Symposium on Advances in Image and Video Technology*, Springer-Verlag, 2008, S. 611–623

[Koch u. Yang 1998] KOCH, K.R. ; YANG, Y.: Robust Kalman filter for rank deficient observation models. In: *Journal of Geodesy* 72 (1998), August, S. 436–441

[Kolmogoroff 1933] KOLMOGOROFF, A.N.: *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933 ( 104). – 415–458 S.

[Kolmogorov u. Zabih 2001] KOLMOGOROV, V. ; ZABIH, R.: Computing Visual Correspondence with Occlusions using Graph Cuts. In: *International Conference on Computer Vision*, 2001, S. 508–515

[Kolmogorov u. Zabih 2004] KOLMOGOROV, V. ; ZABIH, R.: What Energy Functions can be Minimized via Graph Cuts? In: *IEEE Transaction on Pattern Analysis and Machine Intelligence* 26 (2004), February, Nr. 2, S. 147 – 159

[Korte u. Vygen 2008] KORTE, B. (Hrsg.) ; VYGEN, J. (Hrsg.): *Combinatorial Optimization : Theory and Algorithms*. Fourth Edition. Berlin, Heidelberg : Springer-Verlag Berlin Heidelberg, 2008 (Algorithms and Combinatorics 21). – In: Springer-Online

[Krebs 1999] KREBS, Volker: *Nichtlineare Filterung*. Universität Karlsruhe (TH) : Institut für Regelungs- und Steuerungssysteme, 1999

[Labayrade u. a. 2002] LABAYRADE, R. ; AUBERT, D. ; TAREL, J.P.: Real time obstacle detection in stereovision on non flat road geometry through "v-disparity"representation. In: *IEEE Intelligent Vehicle Symposium* Bd. 1, 2002, S. 646–651

[Larsen u. a. 2006] LARSEN, S. ; PHILIPPOS, M. ; POLLEFEYS, M. ; FUCHS, H.: Simplified Belief Propagation for Multiple View Reconstruction. In: *3D Data Processing, Visualization and Transmission*, IEEE Computer Society, 2006, S. 342–349

[Leibe u. a. 2007] LEIBE, B. ; CORNELIS, N. ; CORNELIS, K. ; VAN GOOL, L.: Dynamic 3D Scene Analysis from a Moving Vehicle. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, 1–8

[Li u. Zucker 2005] LI, G. ; ZUCKER, S. W.: Stereo for Slanted Surfaces: First Order Disparities and Normal Consistency. In: *Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2005, S. 617–632

[Li 2009] LI, S.Z. ; SINGH, S. (Hrsg.): *Markov Random Field Modeling in Image Analysis*. 3. Springer-Verlag London, 2009 (Advances in Pattern Recognition)

[Logothetis u. Krishnamurthy 1999] LOGOTHETIS, A. ; KRISHNAMURTHY, V.: Expectation maximization algorithms for MAP estimation of jump Markov linear systems. In: *IEEE Transactions Signal Processing* 47 (1999), Aug, Nr. 8, S. 2139–2156

[Longuet-Higgins 1981] LONGUET-HIGGINS, H.C.: A computer algorithm for reconstructing a scene from two projections. In: *Nature* 293 (1981), S. 133–135

[Longuet-Higgins u. Prazdny 1980] LONGUET-HIGGINS, H.C ; PRAZDNY, K.: The interpretation of a moving retinal image. In: *Proceedings of Royal Society of London* 208 (1980), S. 385–397

[Lowe 2004] LOWE, D.G.: Distinctive image features from scale-invariant keypoints. In: *International Journal of Computer Vision* 60 (2004), Nr. 2, S. 91–110

[Lucas u. Kanade 1981] LUCAS, B.D. ; KANADE, T.: An iterative image registration technique with an application to stereo vision. In: *Proceedings of Imaging understanding workshop*, 1981, S. 121–130

[Lulcheva u. a. 2008] LULCHEVA, I. ; HUMMEL, B. ; BACHMANN, A.: Probabilistisch-logische Objektklassifikation für Verkehrsszenen. In: *Workshop Fahrerassistenzsysteme FAS*, 2008, S. 38–48

[MacLean 1999] MACLEAN, W.J.: Removal of Translation Bias when Using Subspace Methods. In: *International Conference on Computer Vision*, IEEE Computer Society, 1999, S. 753–758

[Mahalanobis 1936] MAHALANOBIS, P.C.: On the generalised distance in statistics. In: *Proceedings National Institute of Science, India* Bd. 2, 1936, 49–55

[Mahler u. Ronald 2007] MAHLER, R. ; RONALD, P.S.: *Statistical Multisource-Multitarget Information Fusion*. Norwood, MA, USA : Artech House, Inc., 2007

[Mandelbaum u. a. 1999] MANDELBAUM, R. ; SALGIAN, G. ; SAWHNEY, H.: Correlation-based estimation of ego motion and structure from motion and stereo. In: *Proceedings of the IEEE International Conference on Computer Vision* Bd. 1, 1999, S. 544–550

[Mansouri u. Konrad 2003] MANSOURI, A.R. ; KONRAD, J.: Multiple motion segmentation with level sets. In: *IEEE Transactions on Image Processing* 12 (2003), February, Nr. 2, S. 201–220

[Marr 1982] MARR, D.: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. 1. W.H. Freeman, 1982

[Marroquin u. a. 1987] MARROQUIN, J. ; MITTER, S. ; POGGIO, T.: Probabilistic Solution of Ill-Posed Problems in Computational Vision. In: *Journal of American Statistical Association* 82 (1987), March, Nr. 397, S. 76–89

[Mellouli u. Suhl 2009] MELLOULI, T. (Hrsg.) ; SUHL, L. (Hrsg.): *Optimierungssysteme : Modelle, Verfahren, Software, Anwendungen*. Berlin, Heidelberg : Springer-Verlag Berlin Heidelberg, 2009 (Springer-Lehrbuch). – In: Springer-Online

[Metropolis u. a. 1953] METROPOLIS, N. ; ROSENBLUTH, A. ; ROSENBLUTH, M. ; TELLER, A. ; TELLER, E.: Equation of State Calculations by Fast Computing Machines. In: *International Journal of Chemical Physics* 21 (1953), S. 1087–1092

[Mohan u. a. 2001] MOHAN, A. ; PAPAGEORGIOU, C. ; POGGIO, T.: Example-Based Object Detection in Images by Components. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001), Nr. 4, S. 349–361

[Molnar u. Modestino 1998] MOLNAR, K.J. ; MODESTINO, J.W.: Application of the EM algorithm for the multitarget/multisensor tracking problem. In: *IEEE Transactions Signal Processing* 46 (1998), Jan, Nr. 1, S. 115–129

[Mueller u. a. 2008] MUELLER, D. ; MEUTER, M. ; PARK, S.B.: Motion Segmentation Using Interest Points. In: *IEEE Intelligent Vehicles Symposium*, 2008, S. 19–25

[Nakayama u. Loomis 1974] NAKAYAMA, K. ; LOOMIS, J.M.: Optical velocity patterns, velocity sensitive neurons, and space perception: A hypothesis. In: *Perception* 3 (1974), S. 63–80

[Nister u. a. 2004] NISTER, D. ; NARODITSKY, O. ; BERGEN, J.: Visual odometry. In: *Proceedings of the International Conference on Computer Vision and Pattern Recognition* Bd. 1, 2004, S. 652–659

[Ohta 2003] OHTA, N.: Motion Parameter Estimation from Optical Flow without Nuisance Parameters. In: *3rd IEEE Workshop on Statistical and Computational Theories of Vision*, 2003

[Palmer 1999] PALMER, S.E.: *Vision Science: Photons to Phenomenology*. Cambridge, MIT Press., 1999

[Papageorgiou u. Poggio 2000] PAPAGEORGIOU, C. ; POGGIO, T.: A trainable System for Object Detectio. In: *International Journal of Computer Vision* 38 (2000), S. 15–33

[Papert 1966] PAPERT, S.: *The Summer Vision Project*. MIT AI Memo 104-1, July 1966

[Poggio u. a. 1985] POGGIO, T. ; TORRE, V. ; KOCH, C.: Computational vision and regularization theory. In: *Nature* 317 (1985), Nr. 26, S. 314–319

[Raj u. Zabih 2005] RAJ, A. ; ZABIH, R.: A Graph Cut Algorithm for Generalized Image Deconvolution. In: *Proceedings of the IEEE International Conference on Computer Vision*, IEEE Computer Society, 2005, S. 1048–1054

[Rauch u. a. 1965] RAUCH, H.E. ; TUNG, F. ; STRIEBEL, C.T.: Maximum Likelihood Estimates of Linear Dynamic Systems. In: *American Institute of Aeronautics and Astronautics* 3 (1965), Nr. 8, S. 1445–1450

[Rodner 2007] RODNER, E.: *Segmentierung mit Graph-Cut-Methoden*, Lehrstuhl für Digitale Bildverarbeitung, Fakultät für Mathematik und Informatik, Friedrich-Schiller-Universität, Jena, Diplomarbeit, 2007

[Rosten u. Drummond 2006] ROSTEN, E. ; DRUMMOND, T.: Machine learning for high-speed corner detection. In: *European Conference on Computer Vision* Bd. 1, 2006, 430–443

[Roy u. Cox 1998] ROY, S. ; COX, I.J.: A Maximum-Flow Formulation of the N-Camera Stereo Correspondence Problem. In: *Proceedings of the International Conference on Computer Vision*, IEEE Computer Society, 1998, S. 492–502

[Scharstein u. Szeliski 2001] SCHARSTEIN, D. ; SZELISKI, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. In: *International Journal of Computer Vision* 47 (2001), S. 7–42

[Schindler 2005] SCHINDLER, K.: Spatially Consistent 3D Motion Segmentation. In: *International Conference on Image Processing*, 2005, S. 409–412

[Schmid 1958] SCHMID, H.: Eine allgemeine analytische Lösung für die Aufgabe der Photogrammetrie. In: *Bildmessung und Luftbildwesen (BuL); Zeitschrift für Photogrammetrie u. Fernerkundung* 2: 1959/1-12 (1958), S. 103–113

[Schoenemann u. Cremers 2006] SCHOENEMANN, T. ; CREMERS, D.: Near Real-time Motion Segmentation using Graph Cuts. In: *Deutsche Arbeitsgemeinschaft für Mustererkennung DAGM e.V.* Bd. 4174. Berlin, Germany : Springer, September 2006 (LNCS), S. 455–464

[Schreer 2005] SCHREER, O.: *Stereoanalyse und Bildsynthese*. Secaucus, NJ, USA : Springer-Verlag New York, Inc., 2005

[Schuermann u. Kreßel 1992] SCHUERMANN, J. ; KRESSEL, U.: Mustererkennung mit statistischen Methoden. / Daimler-Benz AG, Forschungszentrum Ulm, Institut für Informatik. 1992. – Forschungsbericht

[Schwab 2006] SCHWAB, S.: *Ein hierarchischer Kaskadenansatz zu Multiklassifikation mit AdaBoost unter Verwendung von Haar-Merkmalen*, Universität Karlsruhe (TH), Institut für Mess- und Regelungstechnik, 76 131 Karlsruhe, Diplomarbeit, 2006

[Se u. Brady 2002] SE, S. ; BRADY, M.: Ground Plane estimation, error analysis and applications. In: *Robotics and Autonomous Systems* (2002), Nr. 39, S. 59–71

[Shi u. Malik 1998] SHI, J. ; MALIK, J.: Motion Segmentation and Tracking Using Normalized Cuts. In: *Proceedings of the International Conference on Computer Vision*, IEEE Computer Society, 1998, S. 1154–1160

[Shi u. Tomasi 1994] SHI, J. ; TOMASI, C.: Good Features to Track. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 1994, 593–600

[Soquet u. a. 2007] SOQUET, N. ; PERROLLAZ, M. ; LABAYRADE, R. ; AUBERT, D.: Free Space Estimation for Autonomous Navigation, 2007

[Spitzer 1971] SPITZER, F.: Markov Random Fields and Gibbs Ensembles. In: *The American Mathematical Monthly* 78 (1971), Nr. 2, S. 142–154

[Stein u. Shashua 2000] STEIN, G.P. ; SHASHUA, A.: Model-based brightness constraints: on direct estimation of structure and motion. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (2000), September, Nr. 9, S. 992–1015

[Stiller 1994] STILLER, C.: *Modellbasierte Bewegungsschätzung in Bildfolgen*. VDI-Verlag, Düsseldorf, Rheinisch-Westfälische Technische Hochschule Aachen; Informatik/Kommunikationstechnik, VDI Reihe 10; Nr. 320, 1994

[Stiller u. a. 2009] *Kapitel* Maschinelles Sehen. In: STILLER, C. ; BACHMANN, A. ; DUCHOW, C.: *Handbuch Fahrerassistenzsysteme*. Vieweg + Teubner, Wiesbaden, 2009, S. 198–222

[Streit u. Luginbuhl 1995] STREIT, R.L. ; LUGINBUHL, T.E.: Probabilistic Multi-Hypothesis Tracking / Naval Unterwater Systems Center Newport RI. 1995. – Forschungsbericht

[Szeliski u. a. 2008] SZELISKI, R. ; ZABIH, R. ; SCHARSTEIN, D. ; VEKSLER, O. ; KOLMOGOROV, V. ; AGARWALA, A. ; TAPPEN, M. ; ROTHER, C.: A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (2008), June, S. 1068–1080

[Talukder u. Matthies 2004] TALUKDER, A. ; MATTHIES, L.: Real-Time Detection of Moving Objects form Moving vehicles using Dense Stereo and Optical Flow. In: *International Conference on Intelligent Robots and Systems*, 2004

[Ting u. a. 2007] TING, J.A. ; THEODOROU, E. ; SCHAAL, S.: Learning an Outlier-Robust Kalman Filter / Computational Learning & Motor Control Laboratory University of Southern California. 2007 (CLMC Technical Report Number: TR-CLMC-2007-1). – Forschungsbericht

[Torr u. Zisserman 2000] TORR, P.H.S. ; ZISSERMAN, A.: Feature Based Methods for Structure and Motion Estimation. In: *Vision Algorithms: Theory and Practice, number 1883 in LNCS*, Springer-Verlag, 2000, S. 278–295

[Triggs u. a. 2000] TRIGGS, B. ; MCLAUCHLAN, P. ; HARTLEY, R. ; FITZGIBBON, A.: Bundle Adjustment – A Modern Synthesis. In: TRIGGS, B. (Hrsg.) ; ZISSERMAN, A. (Hrsg.) ; SZELISKI, R. (Hrsg.): *Vision Algorithms: Theory and Practice* Bd. 1883, Springer-Verlag, 2000 (Lecture Notes in Computer Science), S. 298–372

[Trucco u. Verri 1998] TRUCCO, E. ; VERRI, A.: *Introductory Techniques for 3-D Computer Vision*. Upper Saddle River, NJ, USA : Prentice Hall PTR, 1998

[Veksler 1999] VEKSLER, O.: *Efficient Graph-Based Energy Minimization Methods in Computer Vision*, Cornell University, New York (USA), Diss., 1999

[Wang u. Adelson 1994] WANG, J.Y.A. ; ADELSON, E.H.: Representing moving images with layers. In: *IEEE Transactions on Image Processing* 3 (1994), September, Nr. 5, S. 625–638

[Wedel u. a. 2009] WEDEL, A. ; MEISSNER, A. ; RABE, C. ; FRANKE, U. ; CREMERS, D.: Detection and Segmentation of Independently Moving Objects from Dense Scene Flow. In: *Proceedings of the International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, Springer-Verlag, 2009, S. 14–27

[Wedel u. a. 2008] WEDEL, A. ; RABE, C. ; VAUDREY, T. ; BROX, T. ; FRANKE, U. ; CREMERS, D.: Efficient Dense Scene Flow from Sparse or Dense Stereo Data. In: *European Conference on Computer Vision*, Springer-Verlag, 2008, S. 739–751

[Wedel u. a. 2007] WEDEL, A. ; SCHOENEMANN, T. ; BROX, T. ; CREMERS, D.: WarpCut - Fast Obstacle Segmentation in Monocular Video. In: *Deutsche Arbeitsgemeinschaft für Mustererkennung DAGM e.V.*, 2007, S. 264–273

[Whittle 1963] WHITTLE, P.: Stochastic process in several dimensions. In: *Bulletin of the International Statistical Institute* 33 (1963), Nr. 40, S. 974–985

[Wishner u. a. 1969] WISHNER, R.P. ; TABACZYNSKI, J.A. ; ATHANS, M.: On the estimation of the state of noisy nonlinear multivariable systems. In: *Automativa* 5 (1969), S. 487–496

[Won u. Derin 1992] WON, S.S. ; DERIN, H.: Unsupervised segmentation of noisy and textured images using Markov random fields. In: *Graphical Models Image Processing* 54 (1992), Nr. 4, S. 308–328

[Younis 1996] YOUNIS, K.S.: *Weighted Mahalanobis Distance for Hyper-Ellipsoidal Clustering*, Air Force Institute of Technology, US, Diplomarbeit, 1996

[Yu u. a. 2006] YU, Y.K. ; WONG, K.H. ; CHANG, M.M.Y. ; OR, S.H.: Recursive Camera-Motion Estimation With the Trifocal Tensor. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics,* 36 (2006), October, Nr. 5, S. 1081–1090

[Yuille 1987] YUILLE, A.L.: Energy function for early vision and analog networks' / MIT, Cambridge, MA. 1987. – Forschungsbericht

[Zabih u. J. Woodfill 1994] ZABIH, R. ; J. WOODFILL, John: Non-parametric Local Transforms for Computing Visual Correspondence. In: *European Conference on Computer Vision*, Springer-Verlag, 1994, S. 151–158

[Zhang 1992] ZHANG, J.: The mean field theory in EM procedures for Markov random fields. In: *IEEE Transactions on Signal Processing* 40 (1992), Oct, Nr. 10, S. 2570–2583

[Zhang u. a. 1994] ZHANG, J. ; MODESTINO, J. W. ; LANGAN, D.A.: Maximum-likelihood parameter estimation for unsupervised stochastic model-based image segmentation. In: *IEEE Transactions on Image Processing* 3 (1994), Nr. 4, S. 404–420

[Zhu u. a. 2006] ZHU, Z. ; OSKIPER, T. ; NARODITSKY, O. ; SAMARASEKERA, S. ; SAWHNEY, H.S. ; KUMAR, R.: An Improved Stereo-Based Visual Odometry System. In: *Proceedings of Workshop on Performance Metrics for Intelligent Systems*, 2006

[Zhuang u. a. 1988] ZHUANG, X. ; HARALICK, R.M. ; ZHAO, Y.: From depth and optical flow to rigid body motion. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 1988, S. 393–397

# Stichwortverzeichnis

Institut für Mess- und Regelungstechnik
Karlsruher Institut für Technologie

Bei der Entwicklung neuer und innovativer Funktionen im Bereich des autonomen Fahrens und der Fahrerassistenzsysteme ist die Aufgabe der Umgebungswahrnehmung von fundamentaler Bedeutung. Mit dem Ziel einer möglichst detaillierten Abbildung der umgebenden Szene nimmt die visuelle Umfelderfassung hier eine zunehmend wichtige Stellung ein. Besonders wünschenswert ist dabei die Szenenrepräsentation in Form einer dichten Segmentierung, welche relevante Verkehrsobjekte vollständig aus der restlichen Umgebung herauslöst und identifiziert.

In dieser Arbeit werden solche relevanten Objekte als räumlich und zeitlich gleichförmig bewegte Gruppierungen von Szenenpunkten beschrieben. Eine Besonderheit des Verfahrens ist die Tatsache, dass die einzelnen Teilaufgaben der dreidimensionalen Rekonstruktion, Bewegungsschätzung und Segmentierung dabei in einem gemeinsamen Modell beschrieben und in verzahnter Reihenfolge gelöst werden. Die einzelnen Modellparameter werden dabei robust durch merkmalsbasierte Verfahren bestimmt, wobei das aktuelle Segmentierergebnis in der Form eines probabilistischen Assoziationsgewichtes mit in den rekursiven Schätzprozess auf der Basis eines Kalman-Glätters integriert wird. Hierdurch kann eine fortlaufende Verbesserung der verkoppelten Schätzgrößen erreicht werden, deren Güte über die Zeit kontinuierlich zunimmt. Die Erwartung der räumlichen und zeitlichen Konsistenz zusammengehöriger Szenenbereiche wird explizit durch das Modell eines Markov-Zufallsfeldes berücksichtigt.

Das Ergebnis der bewegungsbasierten Szenensegmentierung ist eine Menge von Bildbereichen, die vom Menschen eindeutig als unabhängig bewegte Objekte interpretierbar sind. Die Leistungsfähigkeit des Verfahrens wird anhand von realen und synthetischen Bildsequenzen aufgezeigt.

ISSN: 1613-4214
ISBN: 978-3-86644-541-3