# Fusion kombinierter Stereo- und Fokusserien zur 3D-Rekonstruktion

*Ioana Gheţa, Universität Karlsruhe (TH),*
*Michael Heizmann, Jürgen Beyerer, Fraunhofer-Institut für Informations- und Datenverarbeitung IITB, Karlsruhe*

Manuskripteingang: 29. März 2008; zur Veröffentlichung angenommen: 02. Juni 2008

Um 3D-Information aus Bildern auch für Szenen mit schwacher oder periodischer Struktur mit hoher Robustheit zu erzeugen, werden kombinierte Stereo- und Fokusserien fusioniert. Dazu wird eine Kombination von *depth from stereo* und *depth from (de)focus* verwendet. Zur Analyse des Problems wird ein Signalmodell vorgestellt, das die Abbildung durch ein Kamera-Array bei Defokussierung strahlenoptisch beschreibt. Zur Formulierung der Fusionsaufgabe dient ein Energiefunktional, das mittels numerischer Standardverfahren minimiert wird. Die erzielten Ergebnisse der Fusionsmethode an unterschiedlichen Szenen zeigen eine erhebliche Verbesserung der Tiefenschätzung gegenüber Standardverfahren.

**Schlagwörter**: Bildfusion, multivariate Bildserien, 3D-Rekonstruktion, depth from stereo, depth from (de)focus

# Fusion of Combined Stereo and Focus Series for 3D Reconstruction

Combined stereo and focus image series are fused for obtaining robust 3D information. The combination of depth from stereo and depth from (de)focus leads to good results also when dealing with scenes presenting periodical or weak structure. For analyzing the problem in ray optics, a signal model is presented which describes the imaging of a camera array taking into consideration a defocussed projection. The problem is formulated by using energy functionals which are minimized by means of standard algorithms. The results obtained by applying the method to several scenes show considerably improved results compared to the use of standard algorithms.

**Keywords**: Image fusion, multivariate image series, 3D reconstruction, depth from stereo, depth from (de)focus

## 1 Einleitung

Die 3D-Rekonstruktion von Szenen und Objekten spielt für viele Inspektions- und Überwachungsaufgaben eine wichtige Rolle. Eine Möglichkeit zur Gewinnung räumlicher Information ist die Auswertung von Bildserien, bei denen sich der Abstand eines Szenenpunkts vom Aufnahmesystem in eine charakteristische Variation des Bildinhalts abbildet. Zwei prinzipielle Messeffekte sind dabei relevant: Beim Stereoeffekt (*depth from stereo*)

wird die relative Verschiebung des Szenenpunktes erfasst, wenn Bilder von unterschiedlichen Sensorpositionen aufgenommen werden. Verfahren auf der Grundlage des Fokuseffekts beruhen auf der Bestimmung der Schärfe bzw. Unschärfe des abgebildeten Szenenpunkts in den Bildern der Serie (*depth from (de)focus*). Um die Auswertung zu erleichtern, werden dabei in der Praxis meist Bildserien fusioniert, die unter Variation eines einzigen der beiden erwähnten Kameraparameter aufgenommen worden sind.

Obwohl damit bereits mehrere Ansätze und Verfahren für die 3D-Rekonstruktion zur Verfügung stehen, ist bislang keine Lösung bekannt, die universell anwendbar ist. Die bekannten Verfahren haben spezifische Vor- und Nachteile und werden deswegen für bestimmte szenenabhängige Aufgaben eingesetzt. Wenn mehrere Kameras verwendet werden können, ist *depth from stereo* meist das Standardverfahren zur Rekonstruktion von Szenen, die gut erkennbare Strukturen aufweisen. Wenn jedoch keine oder nur schwach sichtbare Strukturen vorhanden sind, liefern solche Verfahren nur unbefriedigende Ergebnisse [3; 4]. Weitere Schwächen treten bei der Inspektion von Objekten auf, die periodische Muster aufweisen.

Um das Ergebnis zu verbessern, ist es sinnvoll, zusätzliche Informationsquellen für die Tiefenbestimmung zu nutzen. Dazu können inhomogene Quellen wie etwa Triangulationssensoren (z. B. Laserscanner) oder Radarsensoren verwendet werden, was jedoch einen erheblichen Aufwand bei der Aufnahme und Registrierung der Entfernungsdaten mit den Bildern erfordert. Eine elegante Alternative dazu stellt die Auswertung von multivariaten Bildserien dar, die durch Variation eines zweiten Parameters entstehen. Die Variation der beiden Aufnahmeparameter kann simultan erfolgen. Man erhält dadurch kombinierte Bildserien, z. B. kombinierte Stereo- und Fokusserien. Zur Bildaufnahme kann ein Kamera-Array eingesetzt werden, bei dem sich die Aufnahmeparameter der Kameras passend variieren lassen. Da bei solchen homogenen Informationsquellen gleichartige Bilddaten vorliegen, ist die Zusammenführung der Tiefeninformation erheblich einfacher als beim Einsatz inhomogener Quellen.

Zum umfassenden Verständnis des Fusionsproblems ist ein Signalmodell der Bildentstehung hilfreich, das die Abbildung der Szene in die einzelnen Bilder der Serie beschreibt, siehe Abschnitt 2. Dabei werden die Positionen der Szenenpunkte im Raum und die bei der unscharfen Abbildung relevanten strahlenoptischen Eigenschaften der Kameraoptiken berücksichtigt. Mit einem derartigen Signalmodell lässt sich die Bildfusion als Spezialfall eines inversen Problems handhaben.

Die direkte Invertierung des Signalmodells ist aufgrund der Schlechtgestelltheit des Problems nicht möglich. Man ist vielmehr gezwungen, durch zusätzliche Annahmen über die Szene und die Abbildung eine Regularisierung des Fusionsproblems vorzunehmen. Dies wird durch Formulierung von Energiefunktionalen erreicht, bei denen die Zusatzannahmen neben den Optimierungskriterien als Energieterme modelliert werden, siehe Abschnitt 3.

# 2 Signalmodell

Das Signalmodell für das gesamte Kamera-Array besteht aus zwei Teilen: Zunächst wird die Abbildung einer Einzelkamera mit den für das Rekonstruktionsproblem relevanten Eigenschaften modelliert. In einem zweiten Schritt wird der Zusammenhang zwischen den Bildern der Serie unter Berücksichtigung des Stereoeffekts und der unterschiedlichen Fokussierungen der Kameras modelliert.

## 2.1 Signalmodell für eine Kamera

Das Signalmodell für eine Einzelkamera hat das Ziel, die Belichtung eines Sensorelements durch die Szene darzustellen. Dazu wird analysiert, welcher Szenenpunkt mit seiner Strahldichte zur Bestrahlung eines Sensorelements beiträgt.

Jedes infinitesimale bildseitige Sensorelement $\delta B(\boldsymbol{\varphi})$ der Kamera (parametriert durch die Polarkoordinaten $\boldsymbol{\varphi} := (\vartheta, \varphi)^{\mathrm{T}}$ mit dem Azimut $\vartheta$ und dem Polwinkel $\varphi$ relativ zum optischen Zentrum $\mathbf{0}$ und zur optischen Achse des Abbildungssystems) erhält seinen gesamten Strahlungsfluss durch das als verlustfrei angenommene Abbildungssystem (idealisiert als dünne Linse mit der Brennweite $f$) von einem infinitesimalen gegenstandsseitigen virtuellen Flächenelement $\delta G(\boldsymbol{\varphi})$ mit dem Flächenverhältnis

$$A(\delta G(\boldsymbol{\varphi})) = \left(\frac{g_0}{b_0}\right)^2 A(\delta B(\boldsymbol{\varphi}))\,.$$

Das Flächenelement $\delta G(\boldsymbol{\varphi})$ befindet sich in der Ebene der fokussierten Abbildung mit Abstand $g_0$ zu $\mathbf{0}$ (d. h. $1/b_0 + 1/g_0 = 1/f$ mit dem Sensorabstand $b_0$ zu $\mathbf{0}$), siehe Bild 1. Zur Vereinfachung wird im Folgenden die Annahme getroffen, dass die Schärfeebene sich zwischen Szene und Kamera befindet; diese Annahme ist für die Anwendbarkeit des Signalmodells nicht zwingend erforderlich.

Vom gegenstandsseitigen virtuellen Flächenelement $\delta G(\boldsymbol{\varphi})$ wird durch das Abbildungssystem nur solches Licht an das bildseitige Sensorelement $\delta B(\boldsymbol{\varphi})$ übertragen, das innerhalb des Raumwinkels $\Omega_G(\boldsymbol{\varphi})$ auf $\delta G(\boldsymbol{\varphi})$ trifft. Die Größe von $\Omega_G(\boldsymbol{\varphi})$ bestimmt sich in Abhängigkeit vom Durchmesser $D$ der Aperturblende, der Gegenstandsweite für eine fokussierte Abbildung $g_0$

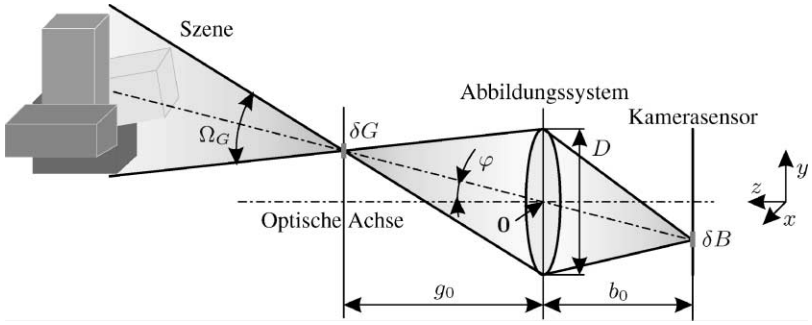


**Bild 1**: Optische Abbildung.
**Figure 1**: Optical projection.

sowie vom Polwinkel $\varphi$ des Flächenelements $\delta G(\boldsymbol{\varphi})$ zur optischen Achse zu

$$\Omega_G(\boldsymbol{\varphi}) = \frac{\pi}{4}\left(\frac{D}{g_0}\right)^2 \cos^3\varphi\,.$$

Den gesamten durch $\delta G(\boldsymbol{\varphi})$ tretenden und auf $\delta B(\boldsymbol{\varphi})$ einfallenden Strahlungsfluss erhält man, indem die Strahldichte $L_G(\boldsymbol{\varphi}, \boldsymbol{\eta})$ über den Raumwinkel $\Omega_G(\boldsymbol{\varphi})$ integriert und mit der Fläche von $\delta G(\boldsymbol{\varphi})$ multipliziert wird:

$$\Phi(\boldsymbol{\varphi}) = A(\delta G(\boldsymbol{\varphi})) \cdot E_G(\boldsymbol{\varphi})$$

$$\text{mit} \quad E_G(\boldsymbol{\varphi}) = \int\limits_{\Omega_G(\boldsymbol{\varphi})} L_G(\boldsymbol{\varphi}, \boldsymbol{\eta}) \cdot \cos\eta \,\mathrm{d}\,\boldsymbol{\eta}\,. \tag{1}$$

Dabei bezeichnen $L_G(\boldsymbol{\varphi}, \boldsymbol{\eta})$ die Strahldichte der vom Flächenelement $\delta G(\boldsymbol{\varphi})$ aus sichtbaren Szenenpunkte in Richtung $\boldsymbol{\eta} := (\xi, \eta)^{\mathrm{T}}$ (parametriert mit dem Azimut $\xi$ und dem Polwinkel $\eta$ zur optischen Achse des Abbildungssystems) und $E_G(\boldsymbol{\varphi})$ die Bestrahlungsstärke von $\delta G(\boldsymbol{\varphi})$.

Die Bestrahlungsstärke des Sensorelements $\delta B(\boldsymbol{\varphi})$ ist somit $E_B(\boldsymbol{\varphi}) = \left(\frac{g_0}{b_0}\right)^2 E_G(\boldsymbol{\varphi})$, aus der sich bei gegebener Belichtungszeit mittels der Koordinatentransformation $\boldsymbol{\varphi} \to \boldsymbol{u}$ mit $\boldsymbol{u}(\boldsymbol{\varphi}) = (u, v)^{\mathrm{T}} = b_0(\tan\varphi\cos\vartheta, \tan\varphi\sin\vartheta)^{\mathrm{T}}$ der dazu proportionale Intensitätswert $B(\boldsymbol{u})$ des Sensorelements $\boldsymbol{u}$ bestimmen lässt.

Befinden sich am Ort des virtuellen Flächenelements $\delta G(\boldsymbol{\varphi})$ nur Szenenpunkte, d. h. findet für $\delta G(\boldsymbol{\varphi})$ eine fokussierte Abbildung statt, und ist die Strahldichte der Szenenpunkte über dem Raumwinkel $\Omega_G(\boldsymbol{\varphi})$ konstant, d. h. $L_G(\boldsymbol{\varphi}, \boldsymbol{\eta}) = L_S(\boldsymbol{x}_G(\boldsymbol{\varphi}), \boldsymbol{\eta}) = L_S(\boldsymbol{x}_G(\boldsymbol{\varphi})) =$ const. für $\boldsymbol{\eta} \in \Omega_G(\boldsymbol{\varphi})$ mit $\boldsymbol{x}_G(\boldsymbol{\varphi}) := \boldsymbol{x}(\boldsymbol{\varphi}; g_0) = g_0 \cdot (\tan\varphi\cos\vartheta, \tan\varphi\sin\vartheta, 1)^{\mathrm{T}}$, so vereinfacht sich die Bestimmung von $E_G(\boldsymbol{\varphi})$ in Gl. (1) erwartungsgemäß [7] zu

$$E_G(\boldsymbol{\varphi}) = \Omega_G(\boldsymbol{\varphi}) \cdot L_S(\boldsymbol{x}_G(\boldsymbol{\varphi})) \cdot \cos\varphi$$

$$\propto L_S(\boldsymbol{x}_G(\boldsymbol{\varphi})) \cdot \cos^4\varphi\,. \tag{2}$$

Zur Bestimmung der jeweils sichtbaren Szenenpunkte und deren Strahldichte $L_G(\boldsymbol{\varphi}, \boldsymbol{\eta})$ muss für jedes gegenstandsseitige Flächenelement $\delta G(\boldsymbol{\varphi})$ der nächstliegende Szenenpunkt $\boldsymbol{x}_S$ in jeder Richtung $\boldsymbol{\eta} \in \Omega_G(\boldsymbol{\varphi})$ z. B. mittels Raytracing ermittelt werden, siehe Bild 2. Am Punkt $\boldsymbol{x}_S$ weist die Oberfläche der Szene eine Strahldichte $L_S(\boldsymbol{x}_S, \boldsymbol{\phi})$ auf, welche das ausgestrahlte Licht in Richtung der lokalen Polarkoordinaten $\boldsymbol{\phi} := (\theta, \phi)^{\mathrm{T}}$ mit dem Azimut $\theta$ und dem Polwinkel $\phi$ zur lokalen Oberflächennormalen beschreibt. Die Strahldichte $L_G(\boldsymbol{\varphi}, \boldsymbol{\eta})$ erhält man demnach aus $L_S(\boldsymbol{x}_S, \boldsymbol{\phi})$ mittels der geometrischen Transformationen $\boldsymbol{x}_S = \boldsymbol{x}_S(\boldsymbol{\varphi}, \boldsymbol{\eta})$ (unter Berücksichtigung der Sichtbarkeit eines Szenenpunkts $\boldsymbol{x}_S$) und $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{x}_S, \boldsymbol{\varphi}, \boldsymbol{\eta})$ zu $L_G(\boldsymbol{\varphi}, \boldsymbol{\eta}) = L_S(\boldsymbol{x}_S(\boldsymbol{\varphi}, \boldsymbol{\eta}), \boldsymbol{\phi}(\boldsymbol{x}_S, \boldsymbol{\varphi}, \boldsymbol{\eta}))$. In diesem Schritt findet also eine Projektion von $\boldsymbol{x}_S \in \mathbb{R}^3$ auf $\boldsymbol{\varphi} \in \mathbb{R}^2$ statt. Die Unschärfe der defokussierten Abbildung wird dabei als Integration der vom Flächenelement $\delta G(\boldsymbol{\varphi})$ aus sichtbaren Szenenpunkte modelliert. Je weiter die Szene von der Ebene der fokussierten Abbildung entfernt ist, desto größer ist – entsprechend der Aufweitung des durch den Raumwinkel $\Omega_G(\boldsymbol{\varphi})$ definierten Kegels – der Einflussbereich eines Szenenpunkts.

Zusammengefasst besteht dieses Signalmodell aus den in Bild 3 dargestellten Komponenten. Ausgehend von der Strahldichte $L_S(\boldsymbol{x}_S, \boldsymbol{\phi})$ der Szene wird in Abhängigkeit von den Parametern $\mathbf{0}$, $f, b_0$ die

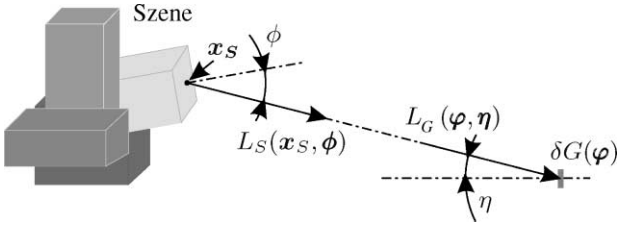


**Bild 2**: Bestimmung der Strahldichte $L_G(\boldsymbol{\varphi}, \boldsymbol{\eta})$.
**Figure 2**: Determination of the radiance $L_G(\boldsymbol{\varphi}, \boldsymbol{\eta})$.

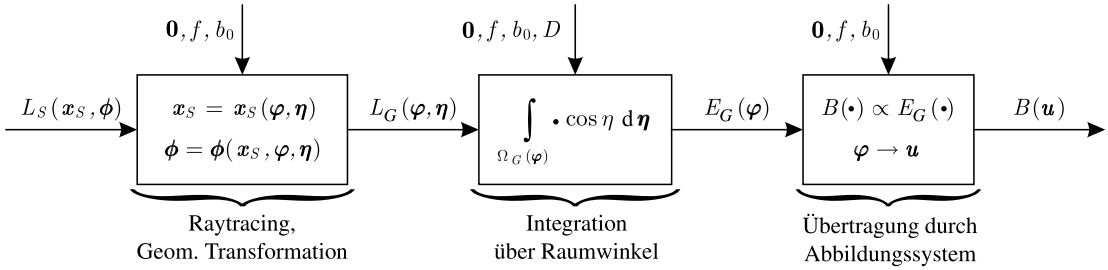


**Bild 3:** Signalmodell für eine Kamera.
**Figure 3:** Signal model of a camera.

Strahldichte $L_G(\boldsymbol{\varphi}, \boldsymbol{\eta})$ am Ort des virtuellen Flächenelements $\delta G(\boldsymbol{\varphi})$ bestimmt. Durch Integration über den Raumwinkel $\Omega_G(\boldsymbol{\varphi})$ mit den Parametern $\mathbf{0}$, $f, b_0, D$ erhält man die Bestrahlungsstärke des Flächenelements $\delta G(\boldsymbol{\varphi})$. Die Übertragung der Bestrahlungsstärke durch das Abbildungssystem mit den Parametern $\mathbf{0}$, $f, b_0$, die zeitliche Integration über die Belichtungszeit sowie die Koordinatentransformation von den Polarkoordinaten $\boldsymbol{\varphi}$ in die üblichen kartesischen Bildkoordinaten $\boldsymbol{u}$ ergeben schließlich das aufgenommene Bild $B(\boldsymbol{u})$.

Im Fall einer Szene, die zumindest stückweise eben und senkrecht zur optischen Achse ist ($z = \Delta g + g_0 =$ const.), lässt sich dieses Signalmodell vereinfachen, siehe Bild 4. Ein Raytracing zur Bestimmung des von jedem Flächenelement $\delta G(\boldsymbol{\varphi})$ aus sichtbaren Szenenbereichs ist nicht notwendig; er ist von vornherein durch die Aufnahmegeometrie bekannt. Für eine kreisförmige Aperturblende mit Durchmesser $D$ ist dies ein kreisförmiger Bereich mit dem konstanten Durchmesser $D_S = \frac{\Delta g}{g_0} D$.

Mit der Annahme, dass die Strahldichte der Szene $L_S(\boldsymbol{x}_S, \boldsymbol{\phi})$ konstant über den jeweiligen Raumwinkel der Betrachtung $\Omega_G(\boldsymbol{\varphi})$ ist, d. h. $L_S(\boldsymbol{x}_S, \boldsymbol{\phi}) = L_S(\boldsymbol{x}_S) =$ const. für $\boldsymbol{\phi} \in \Omega_G(\boldsymbol{\varphi})$, kann die resultierende Bestrahlungsstärke $E_G(\boldsymbol{x}_G(\boldsymbol{\varphi}))$ am virtuellen Flächenelement $\delta G(\boldsymbol{\varphi})$ in kartesischen Koordinaten anstelle Gl. (1) bzw. Gl. (2) als Faltung der Strahldichte der Szene mit einer Impulsfunktion $h(x, y, \Delta g)$ formuliert werden:

$$E_G(\boldsymbol{x}_G(\boldsymbol{\varphi})) = (L_S(\boldsymbol{x}_S) \cdot \cos^4 \varphi) \overset{x,y}{**} h(x, y, \Delta g) \tag{3}$$

$$\text{mit} \quad h(x, y, \Delta g) = \zeta \, \text{rect}\left( \frac{\sqrt{x^2 + y^2}}{\varepsilon(\Delta g)} \right) ,$$

$$\varepsilon(\Delta g) = \frac{D f \Delta g}{(g_0 - f)(g_0 + \Delta g)} ,$$

$$\boldsymbol{x}_S = (g_0 + \Delta g) \cdot (\tan\varphi \cos\vartheta, \tan\varphi \sin\vartheta, 1)^{\mathrm{T}}$$

$$= \frac{g_0 + \Delta g}{g_0} \boldsymbol{x}_G(\boldsymbol{\varphi}) .$$

Dabei sind $\zeta$ eine Konstante mit $[\zeta] =$ sr und $\varepsilon(\Delta g)$ der Durchmesser des Unschärfescheibchens, das zur Belichtung von $\delta G(\boldsymbol{\varphi})$ beiträgt [9].

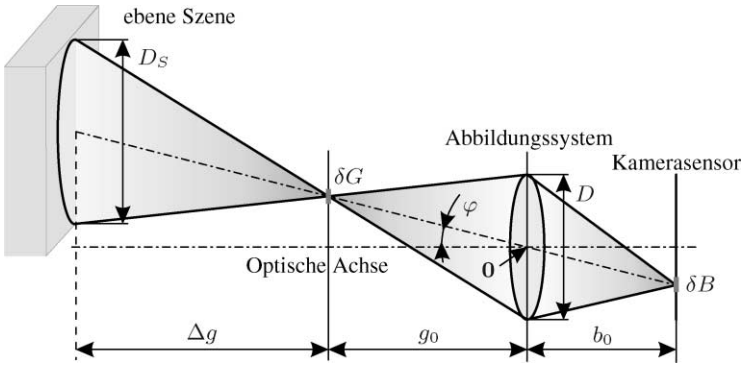


**Bild 4:** Optische Abbildung bei ebener Szene.
**Figure 4:** Optical projection for a planar scene.

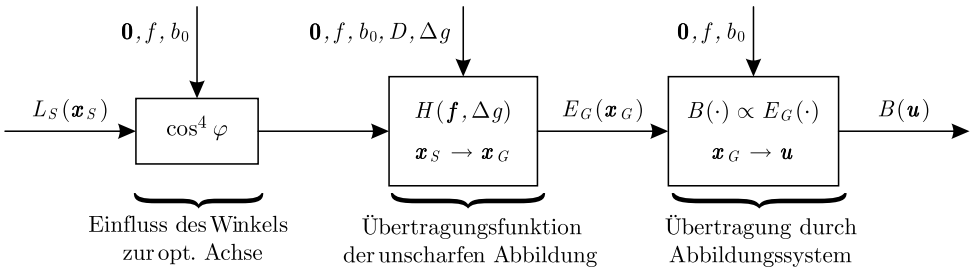


**Bild 5:** Signalmodell für eine Kamera bei ebener Szene.
**Figure 5:** Signal model of a camera for a planar scene.

Bei Übergang in den Frequenzbereich geht die Faltung in Gl. (3) über in eine Multiplikation der Fouriertransformierten:

$$\mathcal{F}\{E(\boldsymbol{x}_G(\boldsymbol{\varphi}))\} = \mathcal{F}\{(L_S(\boldsymbol{x}_S) \cdot \cos^4 \varphi)\} \cdot H(\boldsymbol{f}, \Delta g) \qquad (4)$$

mit $\boldsymbol{f} = (f_x, f_y)^{\mathrm{T}}$. Dabei wird die Übertragungsfunktion $H(\boldsymbol{f}, \Delta g)$ mittels der Besselfunktion erster Gattung, erster Ordnung erhalten:

$$H(\boldsymbol{f}, \Delta g) = \zeta \, \frac{\boldsymbol{J}_1(\pi\varepsilon(\Delta g)||\boldsymbol{f}||)}{||\boldsymbol{f}||} \, . \qquad (5)$$

Die unscharfe Abbildung ist demnach systemtheoretisch eine Tiefpassfilterung mit einer Besselfunktion, wobei die Stärke der Unschärfe vom Abstand der Szene zur Ebene der fokussierten Abbildung nichtlinear abhängt.

Die Bestrahlungsstärke des Sensorelements $\delta B(\boldsymbol{\varphi})$ ist wie oben proportional zur Bestrahlungsstärke von $\delta G(\boldsymbol{\varphi})$ mit $E_B(\boldsymbol{\varphi}) = \left(\frac{g_0}{b_0}\right)^2 E_G(\boldsymbol{x}_G)$. Daraus lässt sich bei gegebener Belichtungszeit mittels der Koordinatentransformationen $\boldsymbol{\varphi} \rightarrow \boldsymbol{u}$ bzw. $\boldsymbol{x}_G \rightarrow \boldsymbol{u} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ mit $\boldsymbol{u}(\boldsymbol{x}_G) = (u, v)^{\mathrm{T}} = \frac{b_0}{g_0}(x_G, y_G)^{\mathrm{T}}$ der dazu proportionale Intensitätswert $B(\boldsymbol{u})$ des Sensorelements $\boldsymbol{u}$ bestimmen.

Gegenüber dem allgemeinen Signalmodell von Bild 3 ergibt sich für ebene Szenen ein modifiziertes Signalmodell, siehe Bild 5. Die Strahldichte $L_S(\boldsymbol{x}_S)$ der Szene wird in Abhängigkeit von den Parametern $\mathbf{0},\ f, b_0$ mit dem Faktor $\cos^4 \varphi$ gewichtet. Die unscharfe Abbildung wird dann im Frequenzbereich mit der von den Parametern $\mathbf{0},\ f, b_0, D, \Delta g$ beeinflussten Übertragungsfunktion $H(\boldsymbol{f}, \Delta g)$ modelliert, woraus sich mit der Koordinatentransformation $\boldsymbol{x}_S \rightarrow \boldsymbol{x}_G$ die Bestrahlungsstärke $E_G(\boldsymbol{\varphi}, \boldsymbol{\eta})$ am Ort des virtuellen Flächenelements $\delta G(\boldsymbol{\varphi})$ bestimmen lässt. Die Übertragung der Bestrahlungsstärke durch das Abbildungssystem, die zeitliche Integration über die Belichtungszeit und die Bestimmung des aufgenommenen Bildes $B(\boldsymbol{u})$ entsprechen dem allgemeinen Signalmodell, wobei hier die Koordinatentransformation von den Koordinaten $\boldsymbol{x}_G$ in die Bildkoordinaten $\boldsymbol{u}$ erfolgt.

## 2.2 Signalmodell für ein Kamera-Array

Das oben eingeführte Modell für zumindest stückweise ebene Szenen lässt sich zu einem Signalmodell für ein Kamera-Array mit $n$ parallel ausgerichteten Kameras erweitern. Wesentlich hierbei ist zunächst der Stereoeffekt, der implizit zur Erfassung von Stereoserien führt. Wird zusätzlich z. B. die Fokussierung der Einzelkameras variiert, erhält man kombinierte Serien.

Vereinfacht man das Signalmodell für ebene Szenen von Bild 5, indem die $\cos^4 \varphi$-Gewichtung für kleine Winkel $\varphi$ vernachlässigt und anstelle der Bestrahlungsstärke $E_G(\boldsymbol{x}_G)$ direkt der dazu proportionale Intensitätswert $B(\boldsymbol{u})$ verwendet wird, erhält man das in Bild 6 dargestellte Signalmodell für die Abbildung der Strahldichtefunktion der Szene $L_S(\boldsymbol{x}_S, \boldsymbol{\phi})$ auf die $n$ Bilder $B_i(\boldsymbol{u}_i)$. Die Gegenstandsweite geht in Gestalt des Parameters $\Delta g_1$ bzw. der damit zusammenhängenden Parameter $\Delta g_i$ an zwei Stellen in das Signalmodell ein: Zum einen in die Übertragungsfunktion der unscharfen Abbildung $H(\boldsymbol{f}, \Delta g)$ (Fokuseffekt) und zum anderen in die Koordinatentransformationen $\boldsymbol{x}_S \rightarrow \boldsymbol{u}_i$ (Stereoeffekt).

## 2.3 Umkehrung des Signalmodells zur Tiefenrekonstruktion

Das Signalmodell von Bild 6 legt nahe, dass es möglich sein sollte, von einem Bild $B_i(\boldsymbol{u}_i)$ zu einem anderen Bild $B_j(\boldsymbol{u}_j)$ zu gelangen. Dazu müssten die Operatoren im Zweig für $B_i(\boldsymbol{u}_i)$ invertiert werden, siehe Bild 7. Diese Umkehrung ist aus mehreren Gründen, die sich auf die unbekannte Gegenstandsweite zurückführen lassen, theoretisch nicht möglich: Die Besselfunktion $H(\boldsymbol{f}, \Delta g_i)$ zur Unschärfemodellierung weist Nullstellen und Stellen mit sehr geringen Werten im Frequenzspektrum auf, sodass die Invertierung dort nicht möglich bzw. störanfällig ist. Zusätzlich ist die Invertierung der Transformation $\boldsymbol{x}_S \rightarrow \boldsymbol{u}_i : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ schlecht gestellt und muss z. B. durch Zusatzinformation regularisiert werden.

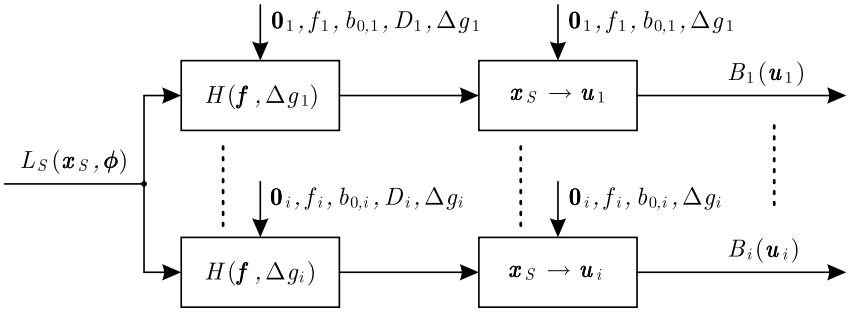


**Bild 6**: Signalmodell für ein Kamera-Array bei ebener Szene.
**Figure 6**: Signal model of a camera array for a planar scene.

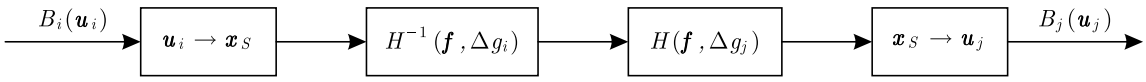


**Bild 7**: Transformationsschritte von $B_i(\boldsymbol{u}_i)$ zu $B_j(\boldsymbol{u}_j)$.
**Figure 7**: Transformation steps from $B_i(\boldsymbol{u}_i)$ to $B_j(\boldsymbol{u}_j)$.

Man kann hingegen versuchen, die Bilder $B_i(\boldsymbol{u}_i)$ und $B_j(\boldsymbol{u}_j)$ möglichst „gut“ aufeinander abzubilden, wobei die Gegenstandsweiten der Szene, d. h. $\Delta g_i$ bzw. $\Delta g_j$, als einzige unbekannte und variierbare Parameter auftreten. Das Problem der Tiefenrekonstruktion wird so zu einem Parameterschätzproblem für die optimalen Gegenstandsweiten $\Delta g_i$ bzw. $\Delta g_j$, sodass die Bilder $B_i(\boldsymbol{u}_i)$ und $B_j(\boldsymbol{u}_j)$ möglichst gut aufeinander abgebildet werden können. Das Gesamtproblem der Bestimmung der Beziehungen zweier Bilder lässt sich dabei auf zwei Teilprobleme reduzieren:

**Koordinatentransformation:** Die Bestimmung der Koordinatentransformation im Signalmodell ist gleichbedeutend mit der Ermittlung der Disparität im Fall von Bildpaaren [10]. Um mit den vorliegenden Bildserien eine möglichst gute 3D-Rekonstruktion zu erzielen, müssen alle Bilder der Serie fusioniert werden [5].

Dazu wird eine Funktion $s$ definiert, die eine Verallgemeinerung der Disparitätsfunktion für rektifizierte Stereopaare ist [2; 6]. Die Funktion $s(\boldsymbol{u}_i)$ weist jedem Bildpunkt $\boldsymbol{u}_i$ im Bild $i$ einen Bezeichner $\alpha \in \mathcal{L}$ für die zugehörige Gegenstandsweite zu:

$$s(\boldsymbol{u}_i) : \mathbb{R}^2 \rightarrow \mathcal{L}\,, \quad s(\boldsymbol{u}_i) = \alpha$$

$$\Leftrightarrow\ \boldsymbol{u}_i \leftrightarrow \boldsymbol{u}_1 \wedge \boldsymbol{u}_1 \leftrightarrow \boldsymbol{u}_2 = (u_1 + \alpha, v_1)^{\mathrm{T}}\,. \tag{6}$$

Alle korrespondierenden Bildpunkte, d. h. alle Bildpunkte, die Abbildung desselben 3D-Punkts der Szene $\boldsymbol{x}_S$ sind, besitzen denselben Bezeichner. Als Bezeichner werden die Disparitäten zwischen einem ausgewählten Referenz-Bildpaar (z. B. den ersten beiden Bildern der Serie) gewählt. Korrespondenzen zwischen Bildpunkten mehrerer Bilder zeigen sich so als identische Werte von $s(\boldsymbol{u}_i)$, d. h. $\boldsymbol{u}_i \leftrightarrow \boldsymbol{u}_j \leftrightarrow \boldsymbol{u}_k \Leftrightarrow s(\boldsymbol{u}_i) = s(\boldsymbol{u}_j) = s(\boldsymbol{u}_k) = \alpha$.

**Berücksichtigung der Unschärfe:** Die defokussierte Abbildung lässt sich durch eine Übertragungsfunktion $H(\boldsymbol{f}, \Delta g)$ modellieren, siehe Abschnitt 2.1. Aus diesem Modell kann die Gegenstandsweite auf zwei Arten bestimmt werden:

Befindet sich die Szene in der Ebene der fokussierten Abbildung, so verschwindet der Abstand $\Delta g$ und die Übertragungsfunktion $H(\boldsymbol{f}, \Delta g)$ nach Gl. (5) wird zu einer Konstanten für alle Ortsfrequenzen: $H(\boldsymbol{f}, \Delta g = 0) = \text{const}$. Andernfalls hat $H(\boldsymbol{f}, \Delta g)$ einen mehr oder weniger stark ausgeprägten Tiefpass-Charakter. Diese Eigenschaft lässt sich durch Merkmale auswerten, welche die Ortsfrequenzen in den Bildern bewerten und dasjenige Bild mit dem höchsten Signalanteil bei hohen Frequenzen bestimmen. Der Ansatz der Maximierung eines solchen Fokusmaßes wird als *depth from focus* bezeichnet. Bei diesem Ansatz spielt die Ortsfrequenz der sichtbaren Textur der Szene eine entscheidende Rolle. Je höher die Ortsfrequenzen der Textur sind, desto genauer kann die Auswertung des Fokusmaßes erfolgen. Für eine gute Tiefenauflösung wird eine Bildserie mit kleinen Abständen der jeweils fokussiert abgebildeten Ebenen benötigt. Von Vorteil ist, dass dichte Tiefenkarten erhalten werden, d. h. für jeden Bildpunkt wird ein Tiefenwert erhalten. Die Qualität der Tiefenschätzung lässt sich für jeden Bildpunkt mit einem Konfidenzmaß bewerten.

Beim Ansatz *depth from defocus* wird die „Verschmierung“ von Bildstrukturen durch die Faltung mit der Impulsfunktion $h(x, y, \Delta g)$ nach Gl. (3) bei der defokussierten Abbildung zur Tiefenbestimmung genutzt. Dazu kann z. B. die Breite von Intensitätsübergängen im Bild an scharfen Intensitätskanten der Szene verwendet werden. Die Auswertung der Unschärfe ist nur an möglichst geraden und scharfen Bildkanten sinnvoll. Die resultierenden Tiefenkaren sind daher nicht dicht und enthalten nur an Objekt- bzw. Texturkanten zuverlässige Tiefenwerte. Von Vorteil ist, dass unter bestimmten Voraussetzungen ein Bild hinreichend für eine Tiefenschätzung ist.

# 3 Energiefunktionale zur Tiefenschätzung

Die im Abschnitt 2.3 dargestellten Ansätze lassen sich mittels Energietermen formulieren. Jeder Energieterm steht für einen Informationsbeitrag, z. B. (De)Fokus- oder Stereoeffekt, und besitzt umso kleinere Werte, je „besser“ das Ergebnis ist. Die Minimierung des aus den Termen zusammengesetzten Energiefunktionals kann mittels Standardverfahren, z. B. Graph-Cuts-Algorithmen [8; 10], erfolgen und liefert als Ergebnis eine verallgemeinerte Disparitätsfunktion $s$ nach Abschnitt 2.3, welche die optimale Tiefenschätzung für die jeweils vorliegende Information repräsentiert.

## 3.1 *Depth from Stereo*

Zur Auswertung reiner Stereoserien werden drei Energieterme definiert [3]:

$$E_{\text{stereo}}(s) = E_{\text{data}}(s) + E_{\text{smoothness}}(s) + E_{\text{visibility}}(s) \,. \qquad (7)$$

Der Datenterm $E_{\text{data}}(s)$ stellt die Photokonsistenz sicher, d. h. korrespondierende Bildpunkte sollen ähnliche Grauwerte aufweisen. Der Glattheitsterm $E_{\text{smoothness}}(s)$ fordert, dass außer an Intensitätskanten keine Disparitätssprünge auftreten. Der Sichtbarkeitsterm $E_{\text{visibility}}(s)$ schließt physikalisch unmögliche geometrische Aufnahmekonfigurationen aus.

## 3.2 Fusion von *Depth from Stereo* und *Depth from Focus*

Zur Auswertung des Fokuseffekts mittels *depth from focus* ist üblicherweise eine Bildserie mit derselben Perspektive notwendig (kollokierte Bildsensoren). Kombinierte Stereo- und Fokusserien, die mittels eines Kamera-Arrays aufgenommen worden sind, müssen daher zunächst z. B. mittels Image-Warping-Verfahren unter Berücksichtigung der entsprechenden Disparitäten lokal geometrisch transformiert werden [1].

Das entwickelte Verfahren schätzt die Gegenstandsweite für jeden Bildpunkt auf zwei Arten: erstens durch Stereoauswertung aus (näherungsweise) fokussierten Bildern und zweitens mittels Fokusauswertung der Fokusserie. Da sowohl Stereo- als auch Fokusauswertung dichte Tiefenkarten erzeugen, tragen die entsprechenden Energieterme für das gesamte Bild zur Tiefenschätzung bei. Der Unterschied der Tiefenschätzungen wird mittels eines Konfidenzmaßes bewertet und in das Energiefunktional von Gl. (7) integriert [3]:

$$E_{\text{fusion},1}(s) = E_{\text{stereo}}(s) + E_{\text{focus}}(s) \,.$$

## 3.3 Fusion von *Depth from Stereo* und *Depth from Defocus*

Die Auswertung des Fokuseffekts mittels *depth from defocus* kann wie oben erläutert nur an Kanten erfolgen. Dafür wird in mehreren defokussierten Bildern die Gegenstandsweite an möglichst geraden Kanten geschätzt. Durch Image-Warping ist deren Position auch in den fokussierten Bildern bekannt, sodass die Tiefenschätzung aus *depth from defocus* zur Verbesserung der aus *depth from stereo* gewonnenen Tiefenschätzung in der Umgebung der Kanten verwendet werden kann. In das Energiefunktional von Gl. (7) wird dazu ein zusätzlicher Term eingefügt, der die zwei Tiefenschätzungen in der Umgebung von Kanten fusioniert [4]:

$$E_{\text{fusion},2}(s) = E_{\text{stereo}}(s) + E_{\text{defocus}}(s) \,.$$

# 4 Experimentelle Ergebnisse

Die im Abschnitt 3 beschriebenen Verfahren wurden für unterschiedliche Szenen, bei denen übliche Stereoverfahren fehlerhafte Ergebnisse liefern, getestet. Es lässt sich feststellen, dass bei Szenen mit wenig oder periodischer Struktur eine deutliche Verbesserung der Rekonstruktionsergebnisse erzielt werden kann. Dies entspricht der Erwartung an derartige Verfahren, die Zusatzinformation in die Bildfusion einbringen, um für Stereoverfahren problematische Szenen zu rekonstruieren.

Bild 8 links zeigt die kombinierte Stereo- und Fokusserie einer Szene, deren Hintergrund sehr wenig strukturiert ist. Wird die Tiefe nur mittels Stereoauswertung geschätzt, siehe Bild 9(a), weist die Tiefenkarte in solchen Regionen fehlerhafte Tiefenwerte auf (im Bei-

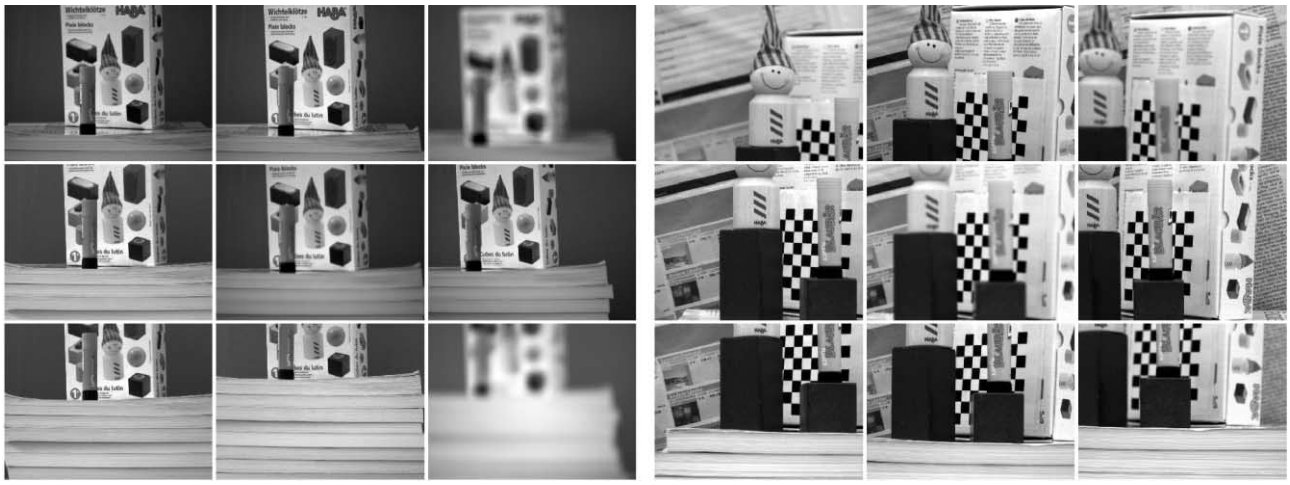

**Bild 8**: Kombinierte Stereo- und Fokusserien zweier Szenen.
**Figure 8**: Combined stereo and focus series of two scenes.

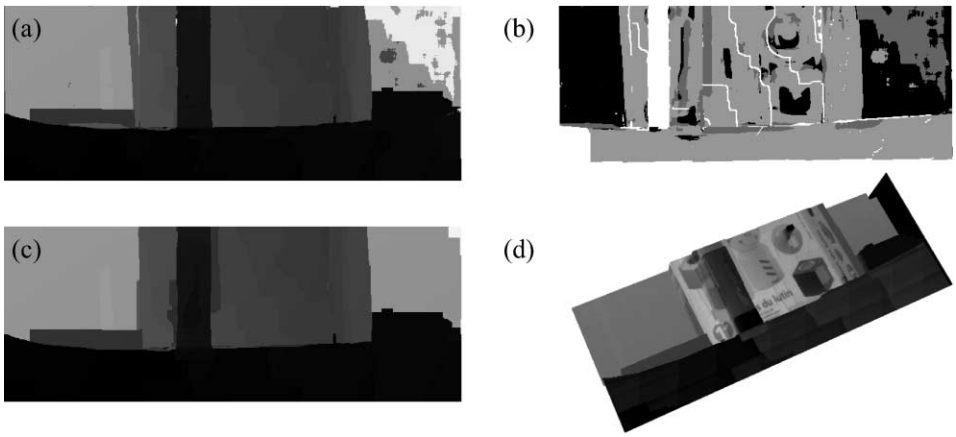


**Bild 9**: Tiefenschätzung aus der Bildserie von Bild 8 links:
(a) Tiefenkarte aus *depth from stereo* (dunkel: nah, hell: entfernt);
(b) Konfidenz der Tiefenkarte (a): schwarz = niedrige Konsistenz, hell grau = hohe Konfidenz, weiß = Verdeckungen;
(c) Verbesserte Tiefenkarte aus der Fusion von Stereo- und Fokusinformation (dunkel: nah, hell: entfernt);
(d) 2 1/2 D-Rekonstruktion der Tiefenkarte (c).
**Figure 9**: Depth estimation from the series in Fig. 8 left:
(a) depth map obtained by *depth from stereo* (dark: near, light: far);
(b) confidence of the depth map (a): black = low confidence, light gray = high confidence, white = occlusion;
(c) improved depth map by fusing stereo and focus information (dark: near, light: far);
(d) 2 1/2 D reconstruction of the scene.

spiel ca. 15 cm Abweichung). Die Konfidenzkarte, siehe Bild 9(b), zeigt in solchen Gebieten durch schwarze Färbung niedrige Konsistenz an. Die Fusion mit der Information aus *depth from focus* verbessert die Tiefenschätzung deutlich, siehe Bild 9(c). Bild 9(d) zeigt die zugehörige 2 1/2 D-Rekonstruktion der Szene.

Das zweite Beispiel in Bild 8 rechts zeigt eine Szene mit periodischem Schachbrettmuster im Hintergrund. Wird nur die Stereoauswertung zur Tiefenschätzung eingesetzt, so weist die Tiefenkarte Fehler im Hintergrund wegen falscher Zuordnung der Bildpunkte in den Bildern der Serie auf (schwarze Streifen), siehe

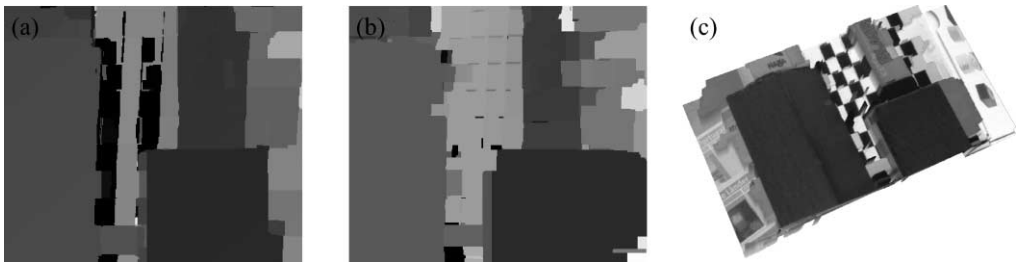


**Bild 10**: Tiefenschätzung aus der Bildserie von Bild 8 rechts:
(a) Tiefenkarte aus *depth from stereo* (dunkel: nah, hell: entfernt);
(b) Verbesserte Tiefenkarte aus der Fusion von Stereo- und Defokusinformation (dunkel: nah, hell: entfernt);
(c) 2 1/2 D-Rekonstruktion der Tiefenkarte (b).
**Figure 10**: Depth estimation from the series in Fig. 8 right:
(a) depth map obtained by *depth from stereo* (dark: near, light: far);
(b) improved depth map by fusing stereo and defocus information (dark: near, light: far);
(c) 2 1/2 D reconstruction of the scene.

Bild 10(a). Das Ergebnis lässt sich wesentlich verbessern, indem die Gegenstandsweite an Kanten mittels der Information aus *depth from defocus* geschätzt und mit der aus der Stereoauswertung erhaltenen Tiefenkarte fusioniert wird, siehe Bild 10(b). Bild 10(c) zeigt schließlich die zugehörige 2 1/2 D-Rekonstruktion der Szene.

# 5 Zusammenfassung

Zur Ausnutzung aller strahlenoptischen Effekte, die einen Rückschluss auf die 3D-Gestalt einer Szene erlauben, ist ein umfassendes Verständnis des Abbildungsprozesses notwendig. Der vorliegende Beitrag beschreibt zunächst die Abbildung einer Szene auf die Bildebene einer Kamera unter der Berücksichtigung einer defokussierten Abbildung. Dieses Signalmodell lässt sich anschließend leicht für ein Kamera-Array erweitern, sodass die wesentlichen Aspekte der Bildentstehung für kombinierte Stereo- und Fokusserien enthalten sind.

Ein Beispiel der Anwendung dieses Signalmodells ist die Fusion kombinierter Stereo- und Fokusserien zur Erzeugung robuster Tiefenkarten für Szenen, die keine oder periodische Struktur aufweisen. In solchen Fällen liefert die Auswertung des Stereoeffekts allein keine zuverlässige Tiefenschätzung. Zusatzinformation kann eingebracht werden, indem die Kameras des Kamera-Arrays unterschiedlich fokussiert werden. Die teilweise unscharfe Abbildung in der Bildserie wird dann mittels Fokus- oder Defokusverfahren ausgewertet und in ein Stereoverfahren integriert. Mittels eines solchen diversitären Prinzips wird eine erhebliche Verbesserung der Tiefenschätzung bei solchen Szenen erzielt.

Darüber hinaus kann das Signalmodell auch für Variationen anderer optischer Parameter erweitert werden, um z. B. Spektral- oder Polarisationsserien zu modellieren. Ein derartiges Modell liefert die Grundlage, um Szenen nicht nur räumlich zu rekonstruieren, sondern zusätzlich auch deren Spektral- bzw. Polarisationseigenschaften zu erkennen.

## Literatur

[1] *M. Bleyer, M. Gelautz:* A layered stereo algorithm using image segmentation and global visibility constraints. In: Proceedings of the 2004 International Conference on Image Processing, S. 2997–3000, 2004.
[2] *O. Faugeras, Q.-T. Luong:* The Geometry of Multiple Images. MIT Press, 2001.
[3] *C. Frese, I. Gheţa:* Robust Depth Estimation by Fusion of Stereo and Focus Series Gained with a Camera Array. In: Proceedings of the IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems, S. 243–248, 2006.
[4] *I. Gheţa, C. Frese, M. Heizmann, J. Beyerer:* A New Approach for Estimating Depth by Fusing Stereo and Defocus Information. In: INFORMATIK 2007, Informatik trifft Logistik, 37. Jahrestagung der Gesellschaft für Informatik e. V. (GI), R. Koschke, O. Herzog, K.-H. Rödiger, M. Ronthaler (Hrsg.), Bd. 1, S. 26–31, 2007.
[5] *I. Gheţa, C. Frese, W. Krüger, G. Saur, N. Heinze, M. Heizmann, J. Beyerer:* Depth Estimation from Flight Image Series Using Multi-View Along-Track Stereo. In: Optical 3D Measurement Techniques, A. Grün, H. Kahmen (Hrsg.), Bd. 2, S. 119–125, 2007.

[6] *R. Hartley, A. Zisserman:* Multiple View Geometry in Computer Vision. 2. Aufl., Cambridge University Press, 2003.
[7] *B. K. P. Horn:* Robot Vision. MIT Press, 1986.
[8] *V. Kolmogorov, R. Zabih:* Multi-camera Scene Reconstruction via Graph Cuts. In: Proceedings of the 7th European Conference on Computer Vision, S. 82–96, 2002.
[9] *F. Mesch:* Systemtheoretische Beschreibung optisch-elektrischer Messsysteme. In: tm – Technisches Messen 7 (8), S. 249–258, 1977.
[10] *D. Scharstein, R. Szeliski:* A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. In: International Journal of Computer Vision 47, S. 7–42, 2002.

1

2

3

**1 Dipl.-Inform. Ioana Gheţa** ist wissenschaftliche Angestellte am Lehrstuhl für Interaktive Echtzeitsysteme, Fakultät für Informatik, Universität Karlsruhe (TH).
**Hauptarbeitsgebiete**: Bild- und Informationsfusion, Mustererkennung, Automatische Sichtprüfung und Bildverarbeitung.
**Adresse**: Institut für Technische Informatik, Lehrstuhl für Interaktive Echtzeitsysteme, Universität Karlsruhe (TH), Adenauerring 4, 76131 Karlsruhe,
E-Mail: ioana.gheta@ies.uni-karlsruhe.de

**2 Dr.-Ing. Michael Heizmann** leitet die Gruppe Variable Bildgewinnung und -verarbeitung am Fraunhofer-Institut für Informations- und Datenverarbeitung IITB.
**Hauptarbeitsgebiete**: Automatische Sichtprüfung und Bildverarbeitung, Bild- und Informationsfusion, Automatisierungstechnik.
**Adresse**: Fraunhofer-Institut für Informations- und Datenverarbeitung IITB, Fraunhoferstraße 1, 76131 Karlsruhe, E-Mail: michael.heizmann@iitb.fraunhofer.de

**3 Prof. Dr.-Ing. Jürgen Beyerer** ist Leiter des Fraunhofer-Instituts für Informations- und Datenverarbeitung IITB und Inhaber des Lehrstuhls für Interaktive Echtzeitsysteme, Fakultät für Informatik, Universität Karlsruhe (TH).
**Hauptarbeitsgebiete**: Automatische Sichtprüfung und Bildverarbeitung, Fusion heterogener Informationsquellen, Informationstheorie, Systemtheorie, Statistische Verfahren, Messtechnik.
**Adresse**: Fraunhofer-Institut für Informations- und Datenverarbeitung IITB und Institut für Technische Informatik, Lehrstuhl für Interaktive Echtzeitsysteme, Universität Karlsruhe (TH), Fraunhoferstraße 1, 76131 Karlsruhe,
E-Mail: juergen.beyerer@iitb.fraunhofer.de