DISS. ETH NO. 22721

# ROBUST AND ACCURATE 3D MOTION ESTIMATION UNDER ADVERSE CONDITIONS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZÜRICH

(Dr. sc. ETH Zürich)

presented by

Christoph Vogel

Dipl. Inform., RWTH Aachen

born on 13.11.1978

citizen of Germany

accepted on the recommendation of

Prof. Dr. Konrad Schindler, examiner
ETH Zürich, Schweiz

Prof. Stefan Roth, Ph.D., co-examiner
TU Darmstadt, Deutschland

Prof. Dr. Daniel Cremers, co-examiner
TU München, Deutschland

2015

# Abstract

3D vision from two eyes and the capability to detect moving objects and estimate their location and movement speed are the basic evidence for humans and animals to interpret and interact with their environment. Likewise, these cues are important for machines, which, equipped with photographic cameras, perceive the world in a similar manner. In the future, intelligent perceiving systems could become omnipresent in our everyday life. Thus, it is not surprising that both, stereo and motion estimation, were in the focus of computer vision since its early days and recently, because of their importance for several key applications like driver assistance, autonomous driving or scene understanding, have regained attention. Somewhat surprisingly, though, both tasks are traditionally tackled individually by the vision community.

This thesis considers the estimation of *scene flow*, the joint computation of motion and geometry from images of (calibrated) stereo cameras, acquired over at least two time steps. Scene flow models appear to have certain advantages over computing geometry and motion individually. Joint inference should allow to exploit correlations between both entities and for instance better reveal co-occurring motion and geometry discontinuities. The capability to lift the motion representation from the image plane to metric 3D space should also be beneficial, as well as the availability of redundant views of the scene.

This work begins with these ideas, and shows that scene flow methods can indeed outperform dedicated stereo and optical methods at their respective task.

In the focus of this thesis are outdoor scenarios, where passive, stereo camera systems possess advantages over active systems like time-of-flight, pattern projection or laser based devices. Under uncontrolled lighting conditions, the constraints imposed by the data can be conflicting or even misleading. Especially in these challenging scenarios, however, scene flow methods can lead to significantly better reconstructions than their 2D counterparts, stereo and optical flow.

We start with a systematic evaluation of different data terms, including several prominent per-pixel and patch-based data costs. Because the data term is vital for high quality reconstructions under adverse conditions, we conduct most of the experiments on a challenging outdoor dataset and try to minimize influences unrelated to the data cost. Motivated by the capability to deliver metric 3D motion estimates, we develop our first scene flow method and focus on the 3D accuracy of the reconstructed motion. Exploiting that many scenes of interest consist of rigidly moving parts, we propose to locally penalize the difference of the flow vectors to a rigid motion. The model prefers locally rigid motions but is not limited to completely rigid scenarios. Building on the experience with this local rigidity assumption we go one step further and

propose to model the scene by planar and over time rigidly moving regions, into which the input images are segmented. Compared to conventional pixel-based representation, significantly less model parameters have to be determined, and jointly estimated along the (over-)segmentation of the scene, leading to accurate geometry and motion boundaries. Finally, the piecewise rigid model is extended by introducing the concept of view-consistency. Here each view holds its own representation of the scene, which are encouraged to be consistent across all frames. This leads to a situation, where all the data of all cameras is treated equally and has to be explained. In practice, view-consistency allows for more efficient occlusion handling and stabilizes the estimation especially in the presence of imaging outliers. Because we employ a scene space parameterization, the model can be easily extended to handle multiple frames in a temporal sliding window, which furthermore increases the redundancy and improves the quality of the reconstruction even further. We evaluate our models on recent datasets, and demonstrate results superior to the state-of-the art for both stereo and motion. Overall the results support the proposition that carefully exploiting the aforementioned advantages in the underlying models, can unveil the potential of scene flow.

# Kurzfassung

Die Fähigkeit des räumlichen Sehens, sowie das Erkennen von Bewegungen in der unmittelbaren Umgebung erlaubt es Menschen und Tieren mit ihrer Umwelt zu interagieren, diese zu interpretieren und zukünftige Aktionen zu planen. Ausgestattet mit photographischen Kameras sind genau diese beiden Qualifikationen oft wesentlicher Bestandteil für Anwendungen, in denen Maschinen mit ihrer Umgebung interagieren und auf Veränderungen reagieren müssen. Gerade solche, intelligenten Maschinen haben das Potential ein fester Bestandteil unserer Zukunft zu werden. Daher ist es wenig überaschend, dass Stereo-Rekonstruktion und das Erfassen von Bewegung aus Kamerabildern, der Optische Fluss, seit jeher ein Kerngebiet der Bildverarbeitung sind. Zudem sind beide Problemstellungen als Schlüsselqualifikation für Fahrerassistenzsysteme und autonomes Fahren auch heute von großer Bedeutung. In der Bildverarbeitung jedoch, werden beide Problem häufig getrennt betrachtet, obwohl Bewegungs- und Tiefen-Information meist gemeinsam benötigt werden.

Diese Arbeit beschäftigt sich mit dem Szenen-Fluss, der gemeinsamen Rekonstruktion von Geometrie und Bewegung aus Bildern kalibrierter Stereo-Kamerasysteme, die in unmittelbarer zeitlicher Folge aufgenommen wurden. Eine gemeinsame Betrachtung beider verwandter Problemstellungen mit Hilfe des Szenen-Flusses erscheint vorteilhaft. Der Szenen-Fluss erlaubt es die Bewegung anstatt in der Bildebene, direkt im metrischen, drei-dimensionalen Raum zu rekonstruieren. Des weiteren sind die Bilddaten in gewisser Redundanz vorhanden und Geometrie und Bewegung korrelierte Einheiten.

An diese Beobachtungen anknüpfend zeigt sich im Verlauf dieser Arbeit, dass mittels Szenen-Fluss gewonnene Rekonstruktionen auch mit spezialisierten Stereo und Optischer-Fluss Methoden mithalten können. Der Fokus dieser Arbeit liegt auf mobilen Anwendungen, in denen keinerlei Kontrolle über die Beleuchtung angenommen werden kann. Gerade für Aussenanwendungen, wie z.B. im Bereich der Fahrassistenz, besitzen passive Stereo-Kamera Systeme Vorteile gegenüber aktiven Methoden wie Time-of-Flight, Laser oder Projektions basierten Messverfahren, oder stellen eine kostengünstige Alternative dar. Ohne Kontrolle über die Beleuchtung können die aufgenommenen Daten allerdings (lokal) häufig irreführend oder widersprüchlich sein und somit die Rekonstruktion erschweren. Eines der Kernargumente dieser Arbeit ist, dass insbesonders unter erschwerten Bedingungen, der Szenen-Fluss der separaten Berechnung von Geometrie und Bewegung mittels Stereo und Optischen Fluss überlegen ist, und zu erheblich besseren Ergebnissen führen kann.

Am Anfang dieser Arbeit werden verschiedene, in der Literatur populäre Datenterme untersucht, wobei der Schwerpunkt auf deren Anwendbarkeit unter widrigen

Bedingungen gelegt ist. Angeregt durch die Fähigkeit zur metrischen Rekonstruktion der Bewegung, liegt der Fokus der ersten entwickelten Methode zur Schätzung des Szenen-Fluss auf deren räumlicher Genauigkeit. Der entwickelte Regularisierer nutzt, dass in vielen Szenarien die Bewegung zumindest lokal als rigide beschrieben werden kann. Der Algorithmus penalisiert dementsprechend lokal die Abweichung der 3D-Flussvektoren zu einer rigiden Bewegung, ist aber nicht nur auf rigide Szenen beschränkt. Der nächste entwickelte Ansatz greift diese Idee auf, und modelliert eine Szene als eine Ansammlung von sich rigide bewegenden Ebenen, in welche die Eingabebilder segmentiert werden. Im Vergleich zu konventionellen, pixelbasierten Ansätzen ist die Anzahl an Modell-Parametern deutlich reduziert. Zudem werden neben den Parametern auch gleichzeitig eine Superpixel-Segmentierung der Szene mitgeschätzt, eine Eigenschaft die zu akkuraten Sprungkanten führen kann. Als letzte Innovation wird das Konzept der Projektions-Konsistenz eingeführt. In diesem Modell hält jede Ansicht der Szene eine eigene Parametrisierung, die aber konsistent gehalten werden. Dadurch sind die Daten aller Kameras gleichberechtigt und müssen durch die Rekonstruktion erklärt werden. In der Praxis erlaubt das Modell ein effiziente Behandlung von Verdeckungen und stabilisiert die Rekonstruktion, im besonderen bei Ausreißern in den Daten. Die Parametrisierung erlaubt dem Modell ebenfalls mehrere Zeitschritte gleichzeitig zu betrachten. Durch die erhöhte Redundanz, verleiht eine Berechnung in einem Zeitfenster dem Algorithmus weitere Stabilität und erhöht die Qualität der Ergebnisse. Eine Auswertung der entwickelten Modelle auf mehreren Datensätzen demonstriert, dass die Verfahren mit an der Spitze der Forschung im Bereich der Bewegungs- und Geometrie-Rekonstruktion aus Stereo-Kameras stehen. Abschließend kann somit eine Bestätigung obiger Thesen gegeben werden. Eine sorgfältige aber konsequente Implementierung der Vorteile des Szenen-Fluss in die Modelle führt hier zu einer höheren Effizienz.