REIHE 12
VERKEHRSTECHNIK/
FAHRZEUGTECHNIK

Fortschritt-Berichte VDI

M.Sc. Malte Oeljeklaus,
Essen

NR. 815

An Integrated Approach for Traffic Scene Understanding from Monocular Cameras

BAND
1|1

VOLUME
1|1

An Integrated Approach for Traffic Scene Understanding from Monocular Cameras

Towards Resource-constrained Perception of Environment Representations with Multi-task Convolutional Neural Networks

DISSERTATION

submitted in partial fulfillment
of the requirements for the degree

Doktor-Ingenieur
(Doctor of Engineering)

in the

Faculty of Electrical Engineering and Information Technology
at TU Dortmund University

by

Malte Oeljeklaus, M.Sc.

Essen, Germany

Date of submission: November 11, 2020

First examiner: Univ.-Prof. Dr.-Ing. Prof. h.c. Dr. h.c. Torsten Bertram
Second examiner: Univ.-Prof. Dr.-Ing. Klaus Dietmayer

Date of approval: May 7, 2021

REIHE 12
VERKEHRSTECHNIK/
FAHRZEUGTECHNIK

Fortschritt-Berichte VDI

M.Sc. Malte Oeljeklaus,
Essen

NR. 815

An Integrated Approach for Traffic Scene Understanding from Monocular Cameras

BAND
1|1

VOLUME
1|1

vdi verlag

https://doi.org/10.51202/9783186815125-I

VDI verlag

Oeljeklaus, M.Sc., Malte
An Integrated Approach for Traffic Scene Understanding from Monocular Cameras
Fortschritt-Berichte VDI, Reihe 12, Nr. 815. Düsseldorf: VDI Verlag 2021.
154 Seiten, 77 Bilder, 24 Tabellen.
ISBN 978-3-18-381512-8, ISSN 0178-9449
57,00 EUR/VDI-Mitgliederpreis: 51,30 EUR

Für die Dokumentation: Szenenverständnis – Umfeldrepräsentation – 3D Rekonstruktion – tiefe neuronale Netze – Multi-task Lernen – geteilte Bildmerkmale – eingebettete Bildverarbeitung – Fortschrittliche Fahrerassistenzsysteme – Automatisiertes Fahren

Keywords: Scene Understanding – Environment Representation – 3D Reconstruction – Convolutional Neural Networks – Multi-task Learning – Feature Sharing – Embedded Computer Vision – Advanced Driver Assistance Systems – Automated Driving

This thesis investigates methods for traffic scene perception with monocular cameras for a basic environment model in the context of automated vehicles. The developed approach is designed with special attention to the computational limitations present in practical systems. For this purpose, three different scene representations are investigated. These consist of the prevalent road topology as the global scene context, the drivable road area and the detection and spatial reconstruction of other road users. An approach is developed that allows for the simultaneous perception of all environment representations based on a multi-task convolutional neural network. The obtained results demonstrate the efficiency of the multi-task approach. In particular, the effects of shareable image features for the perception of the individual scene representations were found to improve the computational performance.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek (German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at www.dnb.de.

D290 (Diss. Technische Universität Dortmund)

ISBN 978-3-18-381512-8, ISSN 0178-9449

Acknowledgement

This thesis was written during my work as a research assistant at the Institute of Control Theory and Systems Engineering of the Faculty of Electrical Engineering and Information Technology of the TU Dortmund University. My special thanks go to Professor Dr.-Ing. Prof. h.c. Dr. h.c. Torsten Bertram for the excellent and ongoing support of my doctoral studies since the early stages until completion and for entrusting me with the scientific freedom that allowed me to write this thesis. Not only through his professional guidance, the constructive discussions and the funding of my doctoral project, but above all through the exceptionally open and trusting atmosphere he created within the team, he undoubtedly played a substantial role in the successful completion of my doctorate. I would also like to thank Professor Dr.-Ing. Klaus Dietmayer for the interest he took in my work, for reviewing this thesis as a second examiner and for the thorough and fair final exam. I would also like to thank Professor Dr.-Ing. Stefan Tappertzhofen for his kind involvement as a third examiner and Professor Dr.-Ing. Martin Pfost for chairing the examination committee.

I thank all employees and former employees of the institute for their friendly and collegial attitude and for their good advice on all scientific and technical problems. In particular, I am thankful to apl. Professor Dr. rer. nat. Frank Hoffmann for his early encouragement and support, which paved the way for me to take up a doctoral project in the first place. I would also like to thank Dr.-Ing. Daniel Schauten for sharing his years of teaching experience and also his experience in working with the university administration.

Furthermore, I want to thank my office colleagues Dr.-Ing. Martin Keller, Dr.-Ing. Javier Antonio Oliva Alonso, Moritz Lütkemöller, Katharina Bartsch, Dr.-Ing. Benedikt Meier and Alexander Hugenroth for enriching my daily work through their constructive and often humorous interaction. Special thanks also go to Dr.-Ing. Christian Wissing, Christian Lienke and Andreas Homann, for the valuable discussions, for sharing their insights on automotive industry practice and also for the great time on our joint conference trip, which I remember very fondly.

I would also like to express my gratitude to Artemi Makarow for his collegial support and especially for his help with the first print version. Likewise, I thank Dr.-Ing. Christoph Rösmann for sharing his extensive experience in software development with me and for establishing modern DevOps tools at the institute, which were extremely helpful for my project. My additional thanks go to Khazar Dargahi Nobari, Faiza Tabassum, and Martin Krüger, with particular thanks for the most memorable scientific and personal conversations during our shared work commutes.

My special thanks as well are due to Jan Braun for his encouragement and mental support during the difficult writing phase. Moreover, I am thankful to my colleagues Katharina Bartsch, Manuel Schmidt, and Jan Braun for their proofreading support and valuable input on the presentation slides. Richard Scherping deserves

my thanks for having initiated my first contact with the institute early on, and also for his friendship since then. In addition, I would also like to thank Franz Albers, Christopher Diehl, Philip Dorpmüller, Robert Gonschorek, Pascal Janke, Maximilian Krämer, Dr.-Ing. Christoph Krimpmann, Dr.-Ing. Jörn Malzahn, Freia Irina Muster, Dr.-Ing. Krishna Kumar Narayanan, Dr.-Ing. Luis Philipe Posada, Niklas Stannartz and Mirko Waldner for the good suggestions and discussions in the doctoral seminars, coffee rounds and lunch breaks.
Naturally, I would also like to mention the technical support for the experimental setups and the IT infrastructure, for which I thank Jürgen Limhoff, Rainer Müller-Burtscheid, Sascha Kersting and Halit Cicek sincerely. The same goes for Gabriele Rebbe and Nicole Czerwinski, whom I thank for their friendly support in all administrative matters.
Some ideas of the present work also arose in the context of student work. For this, I thank the participating students, namely Patrick Weyers, Carlos Miguel Treviño Campa, Marvin Rühl, Björn Polenz, Dizhao Jiang, Zichao Hu and Bernd Möllenbeck.
Gratitude is also owed to all my friends whom I have not mentioned by name for always helping me to find distraction and relaxation when my thoughts were at my work more than due.
Finally, I would like to thank my entire family for their unwavering, unconditional support and encouragement of my personal and professional growth throughout my life. Most importantly, I would like to thank my dear wife Ezgi. Your loving support and patience over the last years were crucial for the completion of this thesis and confirm once more, that as a team we can accomplish just about anything.

Thank you!

Hildesheim, June 2021 Malte Oeljeklaus

Contents

Nomenclature

AOS	average orientation score
α	observation angle
$b, \mathbf{b}, \mathbf{B}$	bias scalar, bias vector, bias tensor
β	position angle
CS	cosine similarity
$\mathrm{concat}(\square)$	concatenation operator, stacks tensors along the u_3 dimension
$\square_{\mathrm{C}}$	marks the use of camera coordinates in [m]
χ	count of frequent classes according to the 85%-15%-rule
$\mathbf{d} = (d_{x_1}, d_{x_2}, d_{x_3})^\top$	vector of 3D bounding box dimensions and its elements in [m]
$\mathrm{diag}(\square)$	diagonal matrix
$\mathcal{E}$	local environment in image space, receptive field
f	focal length of the camera system
$F1$	$F1$-score, harmonic mean of precision and recall
FP	number of false positive samples
FN	number of false negative samples
ϕ	roll angle
$\boldsymbol{\varphi}_{\mathrm{BP}}(\square)$	camera backprojection-line in parametric form
$\varphi_{\mathrm{a}}(\square)$	nonlinear activation function
$\varphi_{\mathrm{s}}(\square)$	softmax function
γ	learning rate for gradient descent optimization
g_{MA}	moving average of the squared gradients
$h, \mathbf{h}, \mathbf{H}$	feature scalar, feature vector, feature tensor (feature map)
$\mathbf{H}^{\diamond}$	feature map with additional rows and columns of zeros
$\mathcal{H}_{\mathrm{Rec}}$	set of features relevant for the topology recognition task
$\mathcal{H}_{\mathrm{Seg}}$	set of features relevant for the road segmentation task
$\mathcal{H}_{\mathrm{Det}}$	set of features relevant for the vehicle detection task
IoU	intersection over union, Jaccard index
IoU_{2D}	2D bounding box IoU in image space
IoU_{BEV}	2D BEV bounding box IoU in world coordinates
IoU_{3D}	volumetric cuboid IoU in world coordinates
i	gradient descent iteration count
$\square_{\mathrm{I}}$	marks the use of image coordinates in [px]
η	road topology class weight
j	general counter index
$\mathbf{K}$	intrinsic camera calibration matrix
κ	class index, discrete category in classification problems
l	indicates the depth of a given neural network layer
$L(\square)$	optimization loss
$L_{\mathrm{L1}}(\square)$	smooth L1 optimization loss

$L_{\mathrm{nll}}(\square)$	negative log likelihood optimization loss, multi-class cross entropy
$L_{\mathrm{Rec}}(\square)$	topology recognition task optimization loss
$L_{\mathrm{Seg}}(\square)$	road segmentation task optimization loss
$L_{\mathrm{Det}}(\square)$	vehicle detection task optimization loss
$L_{\mathrm{2Dbox}}(\square)$	2D bounding box optimization loss
$L_{\varrho}(\square)$	vehicle dimension ratio optimization loss
L_{total}	total optimization loss that combines all perception tasks
$L_{\alpha}(\square)$	observation angle optimization loss
λ_{MA}	moving average decay factor
λ_{LR}	learning rate decay factor
mAP	mean average precision, area under the *pre-rec* curve
$\square_{\mu}$	index denoting a per-sample average, micro average
$\square_{\mathrm{M}}$	index denoting a per-class average, macro average
N_{P}	number of neurons in a neural network layer
N_{L}	number of layers in a neural network
N_{batch}	batch size
N_{Θ}	number of all trainable model parameters
N_{train}	number of samples in the training dataset
N_{test}	number of samples in the test dataset
N_{val}	number of samples in the validation dataset
N_{k}	convolution kernel dimensions
N_{Det}	number of detected bounding boxes
N_{i}	total number of gradient descent iterations
N_{κ}	number of distinguished classes
$\mathbf{n}_{\mathrm{C}}$	scene point position vector in camera coodinates
$\mathbf{n}_{\mathrm{I}}$	scene point position vector in image coodinates
$\mathbf{n}_{\mathrm{W}}$	scene point position vector in world coodinates
$\overset{\infty}{\mathbf{n}}_{\mathrm{W}}$	vanishing point to a given scene point in world coordinates
$\mathbf{n}_{\mathrm{W}}^{C}$	centroid of a 3D bounding box in world coordinates
ν	free parameter of the backprojection line in parametric form
$\mathbf{o}_{\mathrm{C}}$	camera center in camera coodinates in $[\mathrm{m}]$
$\mathbf{o}_{\mathrm{W}}$	camera center in world coodinates in $[\mathrm{m}]$
$\mathbf{o}_{\mathrm{I}} = (o_{u_1}, o_{u_2})^{\top}$	principal point of the camera system
$\mathbf{P} = (\mathbf{p}_1^{\vert}, \mathbf{p}_2^{\vert}, \mathbf{p}_3^{\vert}, \mathbf{p}_4^{\vert}) = (\mathbf{p}_1^{-}, \mathbf{p}_2^{-}, \mathbf{p}_3^{-})^{\top}$	camera projection matrix of size 3×4 and its column and row vectors
pre	precision, positive predictive value
pre_{interp}	interpolated precision
ψ	yaw angle
q	integer multiple of 2π, e.g. $q \in 2\pi \cdot \mathbb{Z}$
$\mathbf{R}$	general rotation matrix of size 3×3
rec	recall, true positive rate
ρ_{κ}	segmentation class a-priori probability

ϱ	aspect ratio of 3D bounding box width and length, e.g. $\varrho = \frac{d_{x_2}}{d_{x_1}}$
ϱ_{NMSE}	normalized mean squared error of the vehicle dimension ratio
s	stride, sliding window step size
TP	number of true positive samples
TN	number of true negative samples
$\mathbf{t}$	general translation vector
τ	decision threshold
θ	general notation of a trainable model parameter
$\Theta = \{\theta_1, \theta_2, \dots \theta_{N_\Theta}\}$	entirety of all trainable model parameters
ϑ	pitch angle
$\tilde{\square}$	accentuation for indicating the use of homogeneous coordinates
$\mathbf{u} = (u_1, u_2, u_3)^\top$	horizontal, vertical and feature channel dimension in image or feature map coordinates
$\mathbf{u}_{\mathrm{mid}} = (u_{1,\mathrm{mid}}, u_{2,\mathrm{mid}})^\top$	2D bounding box midpoint
v	general counter index
$\mathrm{vec}(\square)$	vectorization operator, converts a tensor into a column vector
$w, \mathbf{w}, \mathbf{W}$	weight scalar, weight vector, weight tensor
$\mathbf{W}_{\mathrm{k}}$	convolution kernel
w_{k}	convolution kernel element
w_{Rec}	weight of the topology recognition task in the optimization loss
w_{Seg}	weight of the road segmentation task in the optimization loss
w_{Det}	weight of the vehicle detection task in the optimization loss
$\square_{\mathrm{W}}$	marks the use of world coordinates in [m]
x_1, x_2, x_3	position in spatial world or camera coordinates in [m]
ξ	road topology class occurrence frequency
$y, \mathbf{y}, \mathbf{Y}$	target value, target vector, target tensor
$y_{\mathrm{mid},u_1}, y_{\mathrm{mid},u_2}$	2D bounding box midpoint target variables
$y_{\mathrm{w}}, y_{\mathrm{h}}$	2D bounding box dimensions target variable
ζ	gradient momentum weight factor
$\circ$	Hadamard product, element-wise product
$\lfloor \square \rfloor, \lceil \square \rceil$	floor and ceiling functions, Gauss brackets

Abbreviations and acronyms

ACC	adaptive cruise control
ADAS	advanced driver assistance system
BEV	bird's-eye-view
BL	back left
BR	back right
CNN	convolutional neural network
CPU	central processing unit
CRF	conditional random field
CUDA	compute unified device architecture
FCN	fully convolutional network
FL	front left

FR	front right
GPU	graphics processing unit
ILSVRC	ImageNet large scale visual recognition challenge
InVerSiV	Intelligente Verkehrsinfrastruktur für sicheres vernetztes Fahren in der Megacity
IPM	inverse perspective mapping
KITTI	Karlsruhe Institute of Technology, Toyota Technological Institute
LIDAR	light detection and ranging
LLS	linear least squares
MAC	multiply-accumulate operation
ML	maximum likelihood
MLP	multi-layered perceptron
NAS	neural architecture search
NMSE	normalized mean squared error
px	pixel, image point
RELU	rectified linear unit
SGD	stochastic gradient descent
SSD	single shot detection
YOLO	*here*: you-only-look-once

Abstract

This thesis investigates methods for traffic scene perception with monocular cameras as a foundation for a basic environment model in the context of automated vehicles. The developed approach is designed with special attention to the practical application in two experimental systems, which results in considerable computational limitations. For this purpose, three different scene representations are investigated. These consist of the prevalent road topology as the global scene context and the drivable road area, which are both associated with the static environment. In addition, the detection and spatial reconstruction of other road users is considered to account for the dynamic aspects of the environment. In order to cope with the computational constraints, an approach is developed that allows for the simultaneous perception of all environment representations based on multi-task convolutional neural networks.

For this purpose methods for the respective tasks are first developed independently and adapted to the special conditions of traffic scenes. Here, the recognition of the road topology is realized as general image recognition. Furthermore, the perception of the drivable road area is implemented as image segmentation. To this end, a general image segmentation approach is adapted to improve the incorporation of the a-priori class distribution present in traffic scenes. This is achieved through the inclusion of element-wise weight factors through the Hadamard product, which resulted in increased segmentation performance in the conducted experiments. Also, a task decoder for the perception of vehicles is designed based on a compact 2D bounding box detection method, which is extended by auxiliary regressands. These are used for an appearance-based estimation of the orientation and dimension ratio of detected vehicles. Together with a subsequent method for the reconstruction of spatial object parameters based on constraints derived from the backprojection into the image plane, a scene description with all measurements for a basic environment model and subsequent automated driving functions can be generated. From the examination of alternative multi-task approaches and considering the computational restrictions of the experimental systems, an integrated convolutional neural network architecture is implemented, which combines all perceptual tasks in a single end-to-end trainable model. In addition to the definition of the architecture, a strategy is developed in which alternated training of the perception tasks, changing with each iteration, enables simultaneous learning from several single-task datasets in one optimization process. On this basis, a final experimental evaluation is performed in which a systematic analysis of different task combinations is conducted. The obtained results clearly show the importance of a combined approach to the perception tasks for automotive applications. Thus, the experiments demonstrate that the integrated multi-task architecture for all relevant representations of the scene is indispensable for practical models on realistic embedded processing hardware. Regarding this, especially the existence of common, shareable image features for the perception of the individual scene representations, which are clearly evident from the results, is to be mentioned.

Kurzfassung

Die Arbeit untersucht Wahrnehmungsmethoden mit monokularen Kameras für die Erzeugung eines grundlegenden Umfeldmodells im Kontext automatisierter Fahrzeuge. Der entwickelte Ansatz wird dabei mit Fokus auf die praktische Anwendung in zwei Versuchssystemen ausgelegt, woraus strikte Beschränkungen der rechentechnischen Ressourcen resultieren. Zu diesem Zweck werden drei verschiedene Szenenrepräsentationen untersucht. Diese bestehen aus der Straßentopologie als globalem Szenenkontext und dem befahrbaren Straßenbereich, welche beide dem statischen Umfeld zugerechnet werden. Darüber hinaus wird die Detektion und Rekonstruktion von anderen Verkehrsteilnehmern zur Berücksichtigung der dynamischen Umfeldanteile einbezogen. Um die rechentechnischen Einschränkungen zu berücksichtigen, wird ein Ansatz basierend auf Multi-task Convolutional Neural Networks entwickelt, welcher die gleichzeitige Wahrnehmung aller Umfeldrepräsentationen erlaubt.
Hierzu werden Ansätze für die Wahrnehmungsaufgaben unabhängig voneinander ausgearbeitet und an die Gegebenheiten von Verkehrsszenen angepasst. Die Erkennung der Straßentopologie wird dabei als allgemeine Bilderkennung realisiert. Darüber hinaus wird die Wahrnehmung des befahrbaren Straßenbereichs als Bildsegmentierung umgesetzt. Hierfür wird ein allgemeiner Ansatz zur Bildsegmentierung angepasst um eine stärkere Berücksichtigung der in Verkehrsszenen vorhandenen a-priori Klassenverteilung zu erzielen. Dies erfolgt durch elementweise Gewichtungsfaktoren mittels des Hadamard Produkts, was im Experiment zu einer gesteigerten Segmentierungsgüte führte. Ebenso wird zur Wahrnehmung anderer Fahrzeuge ein Verfahren zur Detektion von 2D Bounding Boxen um zusätzliche Hilfsregressanden erweitert. Diese dienen zur Erscheinungs-basierten Schätzung der Dimensionen sowie der Orientierung detektierter Objekte. Zusammen mit einer Rekonstruktion der räumlichen Parameter durch aus der Rückprojektion in die Bildebene abgeleitete Zwangsbedingungen kann eine für nachfolgende Fahrfunktionen geeignete Objektbeschreibung erzeugt werden. Weiterhin erfolgt, hergeleitet aus der Betrachtung alternativer Multi-task Ansätze und unter Berücksichtigung der rechentechnischen Beschränkungen, die Integration in ein Convolutional Neural Network welches alle Wahrnehmungsaufgaben kombiniert. Zudem wird eine alternierende Trainingsstrategie vorgestellt, welche durch mit jeder Iteration wechselnde Wahrnehmungsaufgaben das simultane Anlernen von mehreren Single-task Datensätzen ermöglicht. Auf dieser Grundlage erfolgt eine abschließende Evaluation, bei welcher eine systematische Untersuchung verschiedener Aufgabenkombinationen erfolgt. Die erzielten Ergebnisse zeigen klar die Bedeutung einer kombinierten Betrachtung der Wahrnehmungsaufgaben für eine Anwendung in der Fahrzeugtechnik auf. So ergibt sich in Hinsicht auf die betrachteten Versuchssysteme, dass eine integrierte Wahrnehmung aller Szenenrepräsentationen für praxistaugliche Modelle unabdingbar ist. In diesem Zusammenhang ist besonders das aus den Ergebnissen ersichtliche Vorhandensein gemeinsamer, mehrfach nutzbarer Bildmerkmale für die Wahrnehmung der einzelnen Szenenrepräsentationen zu nennen.

To my family.