ALGORITHMISCHE BILDRESTAURATION

Algorithmische Kontaminationsbereinigung von **Bilddaten**

Um die Robustheit der optischen Sensorik zu gewährleisten, muss die Informationsgüte der zentralen sensorischen Einheit, der Kamera, sichergestellt werden. In den meisten Fällen ist eine physische Reinigung der Linse aus wirtschaftlichen und technischen Gründen nicht sinnvoll, weshalb digitale Bildrestaurationsalgorithmen entscheidende Vorteile bieten.

Fahrerassistenzsysteme und autonom fahrende Fahrzeuge jeglicher Stufen basieren ihr Perzeptionssystem maßgebend auf optischen Daten, weshalb Kamerasysteme von zentraler Bedeutung für die Autonomous Driving Toolchain sind.

Software-Systeme zur Erkennung von Schildern als auch Verkehrsteilnehmern oder zur semantischen Segmentierung, zum Beispiel von Fahrspuren, verwenden so Bildinformationen um Assistenten wie Emergency Braking, Lane Changing oder Adaptive Cruise Control zu realisieren. Werden diese Informationen nun durch externe Faktoren beeinflusst, wirkt sich dies reziprok auf die Leistung der bildverarbeitenden Algorithmen aus. Diese externen Faktoren äußern sich primär in Kontaminationen der Kameralinse, wie zum Beispiel Verunreinigungen der Kameralinse mit Schmutz, Evaporationsrückstände oder Regen- und Wassertropfen. Somit führt also beispielsweise die Verschmutzung der Frontkamera mit Matsch zu einer kettenreaktionsartigen Performanceverschlechterung des kompletten visuellen Software-Stacks im (teil-)autonomen Fahrzeug, weshalb etwa das Objekterkennungsmodul, bedingt durch den Matsch auf der Linse, nicht mehr dazu in der Lage ist das vorausfahrende Fahrzeug sicher zu identifizieren und daraufhin die sicherheitskritischen Assistenten wie die Active Cruise Control ausfallen können. Um die Robustheit der optischen Sensorik zu gewährleisten, muss also die Informationsgüte der zentralen sensorischen Einheit, der Kamera, sichergestellt werden. In den meisten Fällen ist eine physische Reinigung der Lin-

Bild 1: Akkumulierte Intensitätsdifferenzen nach You et al. [1].

Bild 2: Vergleich zwischen Eingangsbild und Ausgangsbild.

se, wie durch einen Scheibenwischer, aus wirtschaftlichen und technischen Gründen nicht sinnvoll, weshalb digitale Bildrestaurationsalgorithmen entscheidende Vorteile bieten, da keine zusätzlichen Hardwarekomponenten notwendig sind.

Vorverarbeitung und Kontaminationsmaskierung

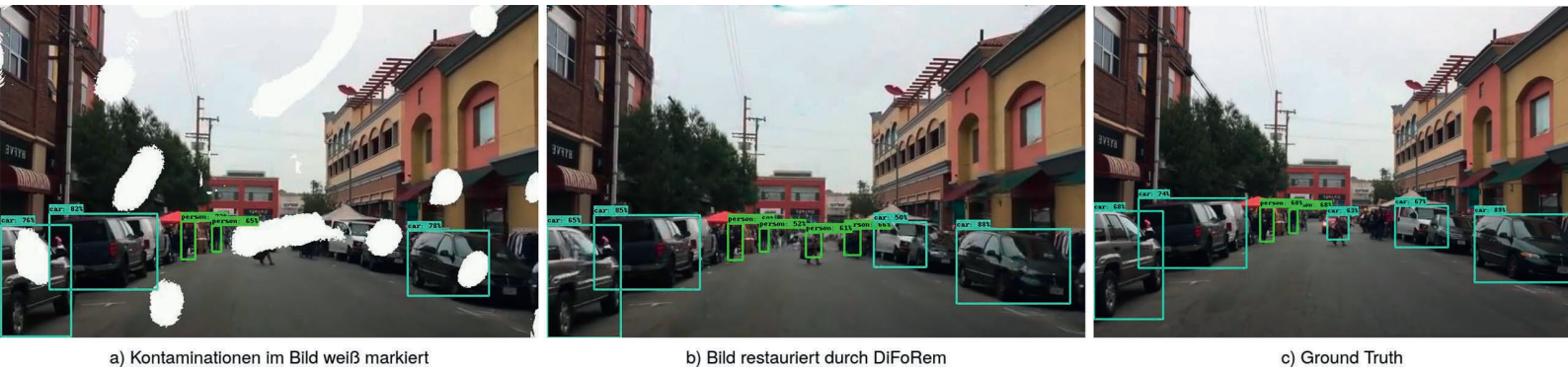
Um Bildverunreinigungen zu entfernen, müssen zunächst Kontaminationen auf Bildebene erkannt und markiert werden. Eine Detektion der kontaminierten Bildbereiche kann über eine Vielzahl an verschiedenen Algorithmen passieren. You et al [1] präsentieren eine Detektionsmethode um Regentropfen anhand temporärer Intensitätsschwankungen im Bild zu detektieren (Bild 1).

Dazu werden Regentropfen nicht als einheitliche Objekte betrachtet, sondern die Erscheinung eines jeden Tropfens wird als kontrahiertes Umgebungsbild, ähnlich dem einer Fish-Eye Kamera, dargestellt. Das Kontraktionsverhältnis eines Wassertropfens liegt bei circa 20 bis 30 [1], was bedeutet, dass Bewegungen, die innerhalb des Tropfens beobachtet werden, 1/20 bis 1/30 langsamer als andere Regionen des Bildes wahrgenommen werden. Jetzt können durch Bewegungsanalyse mittels SIFT-Deskriptoren [2] und durch temporäre Intensitätsschwankungen der Pixelwerte die Merkmale des Bildes extrahiert werden. Anschließend wird der sogenannte SIFT-Flow Algorithmus [3] verwendet um diese SIFT-Deskriptoren zweier oder mehrer konsekutiver Bilder zuzuordnen. Darauffolgend können Regentropfen, nach Rauschreduktion durch Gaussche Faltung, anschließend durch Berechnen der Niveaumengen des resultierenden Bildes identifiziert werden. Mit nachfolgender Feature-Normalisierung und Prüfung der erkannten Regentropfen auf mathematische Geschlossenheit, werden Feature mit Durchmesser kleiner 5 mm entfernt. Schließlich werden die detektierten Regentropfen mittels Pixelmaske maskiert.

Da aber die Kontamination von Kameralinsen durch Regentropfen nur einen kleinen Teil des möglichen Kontaminationsspektrums abdeckt, ist der zuvor beschriebene Algorithmus nicht ausreichend. Daran anknüpfend kann das von Einecke et al. [4] entwickelte Verfahren zur allgemeinen Kontaminationsdetektion abgewandelt werden. Hierfür wird sich ebenso auf die Annahme berufen, dass die Bildposition von möglichen Kontaminationen, zumindest für einen mittel- bis langfristigen Zeitraum, statisch ist. Es werden nur Bilder, die während einer horizontalen Rotation aufgenommen worden sind, verwendet. Bei dieser Rotation um die Y-Achse verändern sich alle Bildkoordinaten unabhängig zum Fokuspunkt, wobei die Koordinaten der kontaminierten Regionen unverändert bleiben. Des Weiteren werden zur Artefaktdetektion, ebenso wie in der zuerst beschriebenen Methode, die zeitlichen Intensitätsdifferenzen der Bildsequenzen verwendet. Da die zeitlichen Intensitätsdifferenzen jedoch sehr sensibel auf wechselnde Lichteinflüsse reagieren, kann stattdessen die normalisierte Kreuzkorrelation (NCC) verwendet werden. NCC ist robust gegenüber schwankender Belichtung [5] und besitzt gleichzeitig eine geringe Laufzeitkomplexität [6]. Mit Hilfe der genannten Maskierungsverfahren können räumliche Informationen über kontaminierte Bildregionen an nachfolgende Restaurationsalgorithmen übermittelt werden.

Bildrestauration und Kontaminationsbereinigung

Da die Rekonstruktion eines komplett kontaminierten Bildes rechnerisch sehr aufwendig ist, wird der Fokus der Rekonstruktion auf die kontaminierten Regionen gelegt. Eine hierzu notwendige Maskierung der zu restaurierenden Bildbereiche wird unter anderem mit den oben genannten Algorithmen bewerkstelligt. Gängige Rekonstruktionsmethoden bekommen also zusätzlich die zu rekonstruierenden Regionen als Eingang.



a) Kontaminationen im Bild weiß markiert

b) Bild restauriert durch DiFoRem

c) Ground Truth

Bild 3: Vergleich der Objektklassifizierungsgüte. © EDAG Engineering

Konventionelle algorithmische Lösungen zur Kontaminationsbereinigung wie zum Beispiel die FastMarching Methode [7] oder Navier-Strokes basierte Methoden [8] haben sich bereits etabliert und sind in diversen Computer Vision Bibliotheken vertreten. Da diese Methoden aber mangels Erfahrungswerten einen sehr naiven Ansatz verfolgen, sind komplexe Strukturen wie zum Beispiel Menschen oder Architektur nur sehr bedingt rekonstruierbar. Das macht diese konventionellen Algorithmen im automobilen Einsatzbereich unbrauchbar.

DiFoRem

Tiefe neuronale Netze lernen, anders wie klassische algorithmische Ansätze, semantische A-priori-Verteilungen und empirisch sinnvolle, versteckte Repräsentationen. Bei kontaminierten Bildregionen zeigen diese Verfahren jedoch Schwächen auf, da sie auf die initialen Werte der kontaminierten Regionen angewiesen sind. Dies zeigt sich häufig in

fehlenden Texturen, starken Kontrastverfälschungen oder zufälligen Kanten in den kontaminierten Bereichen. Um diese Artefakte zu minimieren, können FastMarching Methoden [9] in Kombination mit Poisson Image Blending [10] verwendet werden. Dadurch können jedoch nicht alle Artefakte bereinigt werden [11]. Ferner verwenden diese Methoden regelmäßige, rechteckig maskierte Kontaminationen, welche eine starke Limitierung in Anbetracht des Use-Cases der Kontaminationsbereinigung darstellt.

Partielle Faltungen

Um diesen Einschränkungen entgegen zu wirken, verwendet die von EDAG Engineering präsentierte Lösung „DiFoRem – Real-Time Dirt and Fog Remover" eine spezielle neuronale Netzwerkarchitektur mit partiell faltende Schichten [12]. Diese garantieren, dass bei der Faltung des Bildes nur die kontaminationsfreien Bereiche betrachtet werden und lediglich die maskierten, also die verunreinigten Bereiche, rekonstruiert werden.

Herkömmliche KI-Verfahren zur Bildrestauration verwenden normale faltende neuronale Netze, welche verschmutzte Regionen ebenso betrachten wie saubere. Durch diese Konditionierung des Netzes auf die kontaminierten Bildregionen entstehen häufig Bildartefakte wie Verschwommenheiten und Farbdiskrepanzen, die eine starke Nachverarbeitung zur Beseitigung dieser erfordern. Die von „DiFoRem" verwendete partiell faltende Schicht setzt sich aus drei Schritten zusammen: einer maskierten Faltung und einer renormalisierten Faltung gefolgt von einem Masken-Update Schritt. So wird gewährleistet, dass das Resultat der Faltungsschicht, gegeben eines kontaminierten Eingangsbildes und einer Maske der kontaminierten Regionen, nur auf den kontaminationsfreien Regionen des Bildes beruht und somit agnostisch gegenüber der Art der Kontamination ist. Es wird also eine Konditionierung des Netzes auf Verunreinigungen unterbunden und etwaige visuelle Artefaktbildung präventiv eingedämmt.

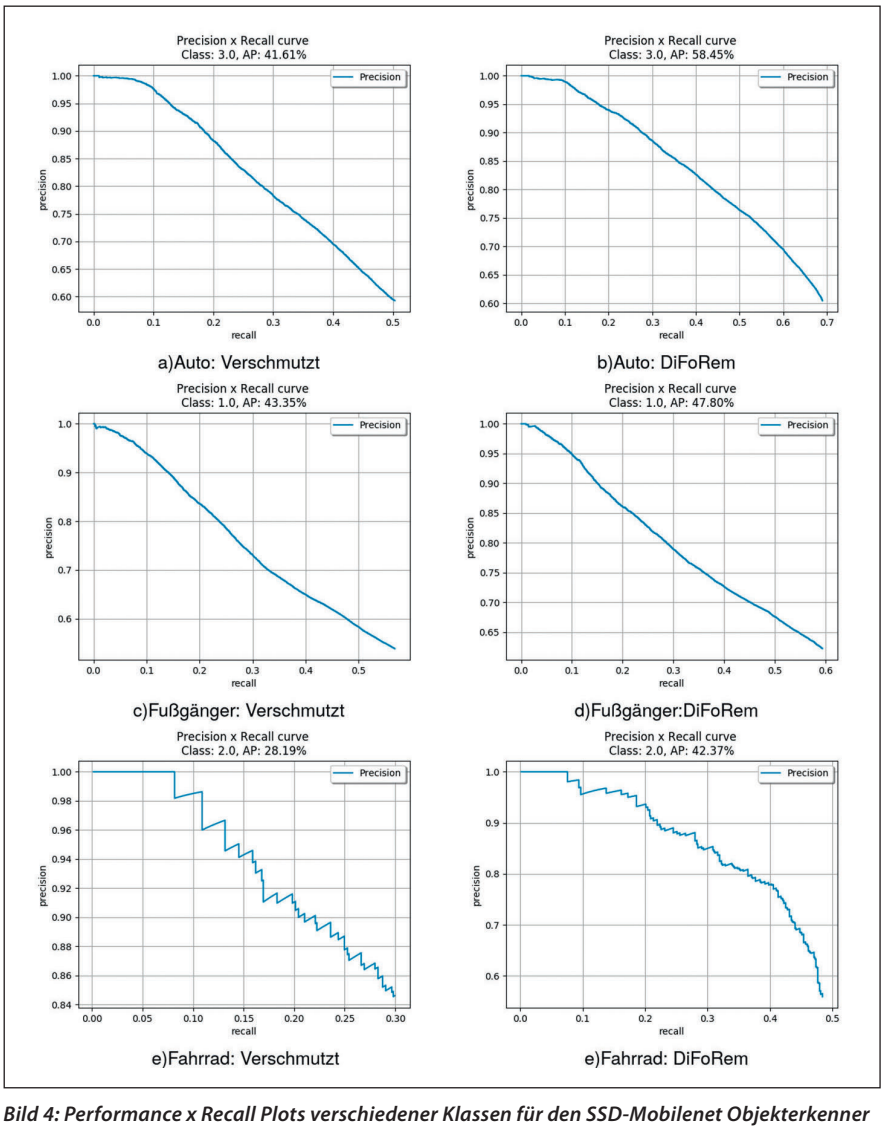


Bild 4: Performance x Recall Plots verschiedener Klassen für den SSD-Mobilenet Objekterkenner jeweils für die verschmutzten und die rekonstruierten Validierungsdaten. © EDAG Engineering

Netzwerkarchitektur

Die Netzarchitektur lehnt sich an die von Ronneberger et al. vorgeschlagene U-Net [13] Architektur an, die sich aus einem Encoder-Decoder-Paar mit sogenannten Skip-connections [14] zwischen den gespiegelten Schichten in Encoder und Decoder zusammensetzt. Diese Encoder-Decoder-Struktur ermöglicht ein effizientes Erlernen von Repräsentationen in Bilddaten, indem im Encoder die Dimensionalität der Eingangsgröße drastisch reduziert wird.

Jeweils der Encoder und der Decoder folgen hier einer typischen CNN-Architektur, wobei die Faltungsschichten durch die oben genannten partiell faltenden Schichten ersetzt werden. Im Decoder wird mittels Nearest-Neighbor Upsampling von der aus der letzten Schicht des Encoders resultierenden latenten Repräsentation das Bild auf Originalgröße rekonstruiert. Die Skip-connections konkatenieren respektive zwei Featuremaps der Eingangsbilder und zwei binäre Kontaminationsmasken und stellen so den Feature- und Maskeninput für die darauffolgende partielle Faltungsschichten dar. So gewährleisten die Skip-connections einen Informationsfluss zwischen Encoder und Decoder und tragen signifikant zur raschen Konvergenz bei. In der letzten partiell faltenden Schicht befindet sich dann das originale kontaminierte Bild konkateniert mit der originalen Kontaminationsmaske. So können Pixel aus kontaminationsfreien Regionen kopiert werden. Um die Informationsgüte des rekonstruierten Bildes auch für Kontaminationen am Bildrand zu gewährleisten, wird Partial Convolution as Padding verwendet, womit vermieden wird, dass der Bildrand durch invalide Werte von außerhalb des eigentlichen Bildes verunreinigt wird.

Loss-Funktion

Um die Konvergenz des Modells sicherzustellen, ist die Loss-Funktion von integralem Bestandteil während des Trainingsprozesses. Da bei Rekonstruktionsaufgaben nicht einfach der mittlere quadratische Fehler zwischen Soll und Ist minimiert werden kann, ist die von EDAG eingesetzte Loss-Funktion von komplexerer Natur. Wird nur auf Pixel-per-Pixel Ebene verglichen, so lernt das Netz nicht die Rekonstruktion von wichtigen High- und Low-Level Features und missachtet somit die Komposition des eigentlichen Eingangsbildes [15]. Sollte jedoch der Pixel-per-Pixel Loss komplett unbetrachtet bleiben, werden große Farb- und Kontrastdifferenzen das rekonstruierte Bild unbrauchbar machen.
Die Loss-Funktion stellt deshalb eine hybride Lösung aus Pixel-per-Pixel und Kompositions-Loss-Funktionen dar. Anfänglich wird der per-pixel Loss, jeweils von den kontaminierten Bereichen und den kontaminationsfreien Bereichen, durch simples Berechnen der mittleren Abweichung bestimmt.
Nun wird der Perceptual-Loss [16] verwendet, wobei ebenso die L1-Distanzen zwischen dem restaurierten Bild, dem restaurierten Bild mit kontaminationsfreien Regionen gleich den Ground-Truth Werten, und dem eigentlichen Ground-Truth Bild berechnet werden. Hierzu werden jedoch nicht einfach die L1-Distanzen der eigentlichen Bilder miteinander verrechnet, sondern das restaurierte Bild.

Um den Fokus nicht zu sehr auf Per-pixel-Differenzen zu legen, wird zusätzlich zu dem Perceptual-Loss noch der Style-Loss berechnet. Dieser wird analog zum Perceptual-Loss ermittelt, jedoch wird vor dem Berechnen der L1-Distanzen eine Autokorrelation der Feature Maps mittels Gram-Matrix durchgeführt. Als letzten Teil der kompletten Loss-Funktion wird der Total-Variation-Loss bestimmt, welcher die smoothing penalty[15] auf eine Region R mit 1-pixel-Dilation um eine kontaminierte Region R^k darstellt.
Als letzten Teil der kompletten Loss-Funktion wird der Total-Variation-Loss bestimmt, welcher die smoothing penalty[15] auf eine Region R mit 1-pixel Dilation um eine kontaminierte Region R_k darstellt.

Rekonstruktions-methode	Verschmut-zungsgrad	Verschmut-zungsgrad
	(0.01, 0.2]	(0.2, 0.4]
$l1_{(PM)}$	**0.78**	2.62
$l1_{(PConv)}$	0.79	2.37
$l1_{(DiFoRem)}$	0.80	**2.34**
PSNR(PM)	30.60	23.20
PSNR(PConv)	31.33	24.02
PSNR(DiFoRem)	**32.78**	**26.34**
SSIM(PM)	0.906	0.721
SSIM(PConv)	**0.907**	0.734
SSIM(DiFoRem)	0.901	**0.747**

Tabelle 1: Vergleich verschiedener Rekonstruktionsmethoden. Spalten stellen prozentuale Verschmutzungsgrade dar.

Mit dem Fokus der Loss-Funktion auf Merkmalswahrung ist „DiFoRem" dazu in der Lage, vielschichtige Bildstrukturen, wie beispielsweise Fassaden, Fahrzeuge oder Personen, mit hoher Genauigkeit zu rekonstruieren. Möglich Artefakte, die es bei per-Pixel basierten Loss-Funktionen gibt, werden somit auch minimiert. Da nicht komplett auf per-Pixel Loss-Funktionen verzichtet wird, können Farb- und Kontrasttreue gesichert werden.

Resultate

Wie in Bild 2 zu sehen ist, kann eine klare Verbesserung der Informationsgüte durch DiFoRem erreicht werden. Da es aber keine fest definierten Metriken zur Evaluation von Rekonstruktionsaufgaben gibt, wird die Performance von DiFoRem analog zu anderen Publikationen im Bereich der Bildrekonstruktion gemessen. Die l1, PSNR und SSIM Metriken wurden auf einem gesonderten, 4000 Bilder großen Test-Datensatz ermittelt. Tabelle 1 vergleicht die State-of-the-Art algorithmische Methode PatchMatch [18] mit der State-of-the-Art Deep Learning Methode PConv [12] und dem von EDAG vorgeschlagenen DiFoRem. Es ist klar erkennbar, dass DiFoRem speziell bei höheren relativen Kontaminationsgraden die anderen Ansätze übertrifft. Besonders bei den Metriken, die die menschliche Wahrnehmung der Rekonstruktionqualität approximieren, wie PSNR und SSIM, überzeugt die EDAG-Methode.

Möchte man nun den Nutzen von DiFoRem für den Softwarestack im Fahrzeug beurteilen, ist die bloße Betrachtung der Bildrestaurationsqualität nicht zielführend. Um diesen Nutzen aufzuzeigen, wird die Performance von Softwaremodulen auf Verbesserung durch Verwenden von DiFoRem überprüft. Exemplarisch zeigt Bild 3 die Performanceverbesserung eines „SSD-MobilenetV1" State-of-the-Art Objektklassifizierers für die Klassen „Fahrradfahrer" und „Auto". Die Performance Metriken wurden wie folgt bestimmt. Zuerst werden die sauberen Validierungsdaten, welche mehrere verschiedene Szenen enthalten, mittels des SSD Mobilenets inferriert, dies stellt nun die Basis des Vergleichs dar. Darauffolgend wird die Sequenz verschmutzt und durch den DiFoRem rekonstruiert. Schließlich werden Objekte in den verschmutzten Bildern und dem von DiFoRem rekonstruierten Bildern ebenfalls mit dem selben SSD Mobilenet erkannt. Nun wird die Precision und der Recall ermittelt und gegeneinander geplottet. Wie in Bild 4 zu sehen ist, ist eine signifikante Verbesserung durch die Rekonstruktion von DiFoRem zu erreichen.

Fazit

Diese Vergleiche verdeutlichen wie einzelne Softwarekomponenten von der erhöhten Informationsgüte, welche DiFoRem gewährleistet, profitieren und so die Gesamtperformance des Perzeptionssystems im Fahrzeug steigern. ■ (oe)

www.edag.com

Das Quellenverzeichnis finden Sie in der Online-Version unter www.hanser-automotive.de

Jacek Burger ist Projektleiter Entwicklungsteam DiFoRem, **Michael Gann** ist im Software Entwicklungsteam DiFoRem und **Lucas Mahler** ebenfalls im Software Entwicklungsteam DiFoRem, alle beschäftigt bei der bei EDAG Electronics, 88131 Lindau.