Shopper Analytics: Ein Ansatz zur Beobachtung von Kundeninteraktionen am Einkaufsregal mit RGB Kameras und Computer Vision

Autor 1: Andreas Gerbl
AI und Marketing
WI7
ag155@hdm-stuttgart.de

Autor 2: Sebastian Sätzler
AI und Marketing
WI7
ss490@hdm-stuttgart.de

Autor 3: Lukas-Orlando Ulmer
AI und Marketing
WI7
lu004@hdm-stuttgart.de

Autor 4: Sania Jasmin Zeidan
AI und Marketing
WI7
sz062@hdm-stuttgart.de

Abstract

Ziel dieses Papers ist die Replikation der Shopper Analytics Studie von Daniele Liciotti et al. [1], mit RGB- Kameras. In Anbetracht der Vorgängerstudie wird das Ziel verfolgt, ein System zu schaffen, welches die Erkennung von Kundenaktivitäten mit einem verteilten RGB-D-Kameranetzwerk ermöglicht.

Unter Anwendung eines hybriden Verfahrens mit tiefen neuronalen Netzen und Homografie sollen Kundeninteraktionen an einem nachgestellten Einzelhandelsregal erfasst werden.

Um dies zu ermöglichen, werden verschiedene Methoden und Konzepte aus dem Bereich computerbasiertes Sehen, wie Deep-Learning und Homografie, implementiert.

Diese Studie setzt sich mit der Auswahl einer geeigneten Softwarearchitektur auseinander, um die Ergebnisse der Originalstudie ohne Tiefensensorik nachzubilden. Hierbei werden Trainingsdaten von drei Produkten erzeugt.

Neben Grundinformationen zur technischen Architektur wird die Einbindung von Testdaten sowie die Darstellung der Ergebnisse visualisiert. Abschließend werden Grenzen der Umsetzung sowie Ansätze zur Optimierung diskutiert.

Keywords: Artificial Intelligence, Object Detection, Homography, Retail, Customer Tracking

1. Einleitung

Im Januar 2018 eröffnete Amazon den ersten Self-Service Store ohne Kassenpersonal oder Bezahlterminals. [2] Kunden der neuartigen Ladens checken sich lediglich per Smartphone in den Store ein, Kameras tracken dann, welche Produkte die Kunden aus dem Regal entnehmen und fügen diese Produkte entsprechend dem virtuellen Warenkorb hinzu. Verlässt der Kunde den Laden, werden die Kosten für den Einkauf direkt vom Amazon Konto abgebucht. Damit bringt Amazon eine Innovation mit Disruptions-Potenzial, denn in Zeiten von einer globalen Pandemie ist ein Retail Konzept, welches ohne Personal und Warteschlangen beim Kassieren daherkommt, äußerst vielversprechend.
Um diese Konzepte umsetzen zu können, werden Systeme benötigt, die Daten erheben und verarbeiten können. Mit deren Hilfe können die Kunden analysiert und "gelesen" werden.
Die Originalstudie "Shopper Analytics: A Customer Activity Recognition System Using a Distributed RGB-D Camera Network" beschreibt einen solchen Aufbau, unter Zuhilfenahme mehrerer RGB Kameras in Kombination mit Tiefensensoren.

Die vorliegende Arbeit ist eine Referenzstudie zu der eben genannten Originalstudie und soll die Ergebnisse ohne Anwendung von Tiefensensoren replizieren. Somit soll eine technisch einfachere Alternative aufgezeigt werden, die nur mit herkömmlichen RGB Kameras umgesetzt werden kann, um eine automatisierte Kundenüberwachung einer breiteren Masse verfügbar zu machen.

Ziel dieser Arbeit ist es, Kundeninteraktionen anhand eines nachgestellten Verkaufsregal für den Einzelhandel mit unterschiedlichen Produkten zu messen und festzuhalten. Untersuchungsobjekt dieses Papers ist ein hybrides Klassifizierungsmodell, welches aus einem tiefen neuronalen Netz sowie einem Homografie-Verfahren zusammengestellt ist, um Kundeninteraktionen zu beobachten und Produkte zu erkennen. Das neuronale Netz übernimmt die Lokalisierung und Erkennung von Händen und Produktkategorien, während die Homografie dazu verwendet wird, das Produkt innerhalb einer Kategorie richtig zuzuordnen.

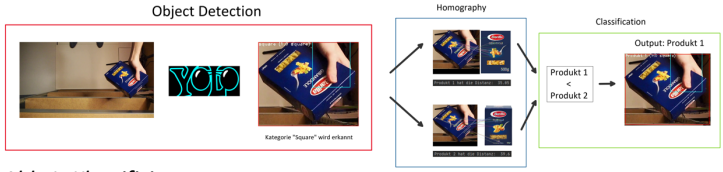

Abb.1: Klassifizierungsprozess

Dabei wurden Kategorien gewählt, die sich optisch stark voneinander unterscheiden, während sich die Produkte innerhalb der jeweiligen Kategorien stark ähneln, um die Homografie auf die Probe zu stellen.

Wie in der Originalstudie werden die Griffe in das Einkaufsregal aufgezeichnet und es wird registriert, wenn Produkte aus dem Regal genommen und wieder zurückgelegt werden.
Dieses Paper beschreibt zunächst den physischen Aufbau des Laborexperiments, gefolgt von einer Einführung in die Objekterkennung und Herleitung der gewählten Objekterkennungsarchitektur.

Wie in der Originalstudie werden die Griffe in das Einkaufsregal aufgezeichnet und es wird registriert, wenn Produkte aus dem Regal genommen und wieder zurückgelegt werden.

Dieses Paper beschreibt zunächst den physischen Aufbau des Laborexperiments, gefolgt von einer Einführung in die Objekterkennung und Herleitung der gewählten Objekterkennungsarchitektur.

2. Methoden

2.1 Technischer Aufbau

Um den rein technischen Aufbau zu beleuchten, wurden anhand des Laborexperiments eigene Trainingsdatensätze aus einer möglichst realitätsnahen Umgebung generiert. Der Aufbauprozess beinhaltet alle visuellen Perspektiven einer Einzelhandelsumgebung und setzt diese analog um.

Zu bemerken ist, dass sich Lichtkonzepte in einem Einzelhandelsgeschäft häufig in zweierlei Arten aufteilen. Transluzente Lichtquellen sorgen für eine allgemein streuende Ausleuchtung des Innenraums, gepaart mit Spotlights, welche eine zielgerichtete Fläche auf den Regal Vorderseiten ausleuchtet. Wichtig bei Spotlights ist es, warmweiße Lichtquellen in einem Lichtfarbenbereich zwischen 2700 bis 3300 Kelvin zu wählen. Dieses Lichtspektrum sorgt für ein wärmeres Licht, wie es beispielsweise in Einkaufsmärkten der Fall ist.

Die Lichtquellen im Laborexperiment waren herkömmlich streuende Halogenlampen, welche für eine indirekte Beleuchtung sorgten. Diese wurden mit einem Studio LED-Ringlicht kombiniert, das sich sowohl im Lichtfarbenbereich als auch in der Ausrichtung einstellen ließ. Die Position des Ringlichtes wurde dabei oberhalb des Regals festgelegt, sodass die Produkte im Regal diagonal von oben beleuchtet wurden.

Die Wahl der Kameraperspektive wurde wie in der Originalstudie die oberhalb des Regals festgesetzt, sodass die Aufnahme lotrecht auf den Boden vor dem Regal gerichtet ist. Dabei bleibt die Regalfront immer noch komplett am Rande des Bildes sichtbar. Ein Unterschied zur Originalstudie besteht darin, dass die Kamera nicht oberhalb des Regalendes befestigt wurde, sondern mit einem Versatz von 12 cm an der oberen Rahmenkante des Regals. Diese Position ist notwendig gewesen, um später zwischen den Höhenabständen einzelner Produkte durch den Größenunterschied der auftauchenden Hand in der Aufnahme deutlich unterscheiden zu können.

In der Einstellung der Kamera mussten auch besondere Vorkehrungen getroffen werden. Eine Hürde beim Erstellen eigener Trainingsdaten in dieser Umgebung ist das Phänomen der Bewegungsunschärfe. Die Betrachtungsgegenstände befinden sich in jeder Aufnahme in Bewegung.

Es wird davon ausgegangen, dass ein Kunde in herkömmlicher Geschwindigkeit Produkte aus dem Regal nimmt. Dabei sind die einzelnen Bilder des Videos zwar scharf gestellt, jedoch ist genau das zu betrachtende Produkt in Bewegung, und es entsteht eine Unschärfe. Vor allem in dieser Konstellation ist diese Unschärfe besonders stark, da die Produkte des oberen Regalfachs eine geringe Distanz zur sich darüber befindenden Kameralinse haben. Dadurch bewegen sich die Produkte durch das Bild schneller und tauchen dabei kürzer auf.

Daher ist eine Verschlusszeit der Kamera zu wählen, deren Dauer mindestens der Anzahl von Bildern pro Sekunde in tausendstel Sekunden beträgt. Im Falle des Laborexperimentes wurde mit 60 Bildern pro Sekunde aufgenommen und eine Verschlusszeit von 1/1250 Sekunden gewählt. Hiermit werden alle diffusen Aufnahmen durch schnelle Bewegungen im Video in jedem einzelnen Bild nahezu restlos aufgelöst.

Eine weitere notwendige Einstellung ist der Bildfokus. Dieser sollte bei einer herkömmlichen Kamera nicht automatisch angepasst werden, da kurzzeitige Unschärfe durch die Neujustierung des Fokus bei Auftreten einer Hand im Bild entsteht.

Demnach sollte der Fokus genau auf das betrachtete Objekt eingestellt sein, wenn es durch das Herausnehmen aus dem Regal erscheint. Für die genannten Einstellungen Beleuchtung und Fokus lassen sich keine pauschalen Richtwerte nennen, da diese je nach Setting und betrachtetem Produkt variieren können.

Sowohl für eine realitätsnahe Umgebung als auch zur Förderung des späteren Trainieren des Modells bietet es sich an, die Aufnahmen auf einem farblich abgegrenzten Boden zu den Produkten vorzunehmen.

Somit ist der Hintergrund der aufgenommenen Produkte weiß, und die Objekte lassen sich besser abgrenzen als auf einem gemusterten Boden. Die Abstände der Regalfächer zueinander betrugen je 40cm. Die Kamera befand sich auf einer Höhe von 180cm. Ein gesamtes Regalfach hatte dabei eine Breite von 90cm. Diese Proportionsverhältnisse sind relevant, da hierbei mit einem Objektiv mit 14mm Brennweite aufgenommen wurde und dennoch auf beiden Seiten des obersten Regalfachs je 5cm an dem Bildrand abgeschnitten wurden.

Dies ist für das Setting jedoch nicht problematisch, da die Objekte nebeneinander mit einem Abstand von 20cm zueinander beim Verlassen des Regals vollkommen sichtbar blieben.

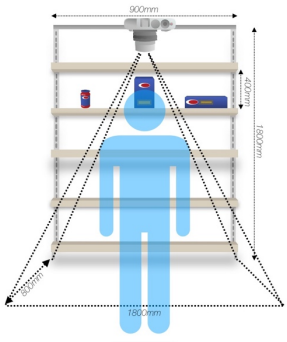

Abb.2: Technischer Aufbau

Für das spätere Labeling lassen sich aus den aufgenommenen Videos einzelne Bilder entnehmen. Dabei werden die Bilder an verschiedenen Stellen beim Durchlaufen des Regalgriffs entnommen.

2.2 Softwaretechnischer Aufbau

2.2.1 Überblick CNN und Objekterkennung

Convolutional Neural Networks (CNNs) sind eine der performantesten Architekturen, um Bildinhalte zu erkennen. IT-Konzerne wie Google und Facebook arbeiten mit eigenen Forschergruppen daran, diese Netzwerke weiter auszubauen und dabei neue, noch bessere CNN Architekturen zu entwickeln. Fast alle neuronalen Netzwerke, die heute zur Bilderkennung verwendet werden, sind CNNs.
Ein Bild besteht aus einem (grayscale) oder mehreren (color) Rastern aus vielen Pixeln. Jedes Pixel enthält Farbinformationen, im RGB Farbraum beispielsweise der Anteil an Rot, Grün oder Blau auf einer Skala von 0-255 (8 Bit) [3].

Die Object Detection (dt. Objekterkennung) ist eine weiterführung des CNN-Ansatzes, die neben dem Label zusätzlich noch Lokalisierungdimensionen ausgibt, in der das Objekt erkannt wird: Die sogenannten Bounding Boxes (dt. Begrenzungsrahmen).
Die Lokalisierungsdimensionen sind zusammengesetzt aus einem Mittelpunkt, sowie einer Höhen- und Weitendimension, welches ein Rechteck um das erkannte Objekt markiert. [4]
Moderne Objekterkennungsarchitekturen besitzen sogenannte Anchor Boxes (dt. Ankerpunkte), welche die Erkennung von mehreren überlappenden Objekten ermöglicht [5]. Dies ist vor allem im Kontext dieser Studie von hoher Bedeutung, da sowohl Hände als auch in der Hand gehaltene Produkte vom Modell getrennt aufgefasst und klassifiziert werden müssen.
Die meisten Object Detection Architekturen sind in 2 Komponenten aufgeteilt: Dem Backbone (dt. Rückgrat) und dem Head (Kopf). Das Rückgrat einer Architektur wird auch als Feature Extractor bezeichnet. Wie aus diesem Begriff zu entnehmen ist, wird das Backbone in der Objekterkennung dazu genutzt, um Merkmale eines Inputs (z.B. eines Bildes) zu extrahieren, welche dem Kopf für die Objekterkennungsoperation verabreicht werden. [6]

Damit Informationen kleiner Objekte im Input-Bild durch häufige Faltungen in tiefen Netzwerken, wie die der Objekterkennung, nicht untergehen, werden Multi-Scale Feature Maps angewandt. [7]

Hierzu werden in den Hidden Layers (versteckte Schichten) frühzeitige Klassifizierungen durchgeführt. Das hat den Effekt, dass die Klassifizierung nicht nur das Ergebnis der letzten versteckten Schicht ist, welche vor allem Merkmale dominanter Objekte eines Inputbildes ausgibt. [8] Damit soll gewährleistet werden, dass Objekte unterschiedlicher Größen zuverlässig klassifiziert werden.

2.2.2 YOLOv4-tiny

Bei der Auswahl des neuronalen Netzes soll zwischen einem Ein-Stufen-Detektor und einem Zwei-Stufen-Detektor entschieden werden. Ein Zwei-Stufen-Detektor durchläuft zwei Prozesse, wobei im ersten Schritt prägnant auffällige Regionen vorhergesagt und anschließend nur in den identifizierten Regionen klassifiziert wird. Bei einem Ein-Stufen-Detektor werden keine Regionen definiert, sondern die Erkennung findet direkt über eine Vorhersage von Klassen und Begrenzungsrahmen in einem Durchlauf statt und läuft folglich schneller ab, wenn auch ungenauer. [6]

Da dieses Experiment die Erkennung von Objekten in Echtzeit vorsieht, wird die Auswahl auf Letzteres fallen.
Wie in der Einleitung angedeutet, wird die Objekterkennung dazu genutzt, übergeordnete Kategorien zu orten. Die Aufgabe der genauen Identifizierung des Produktes wird der Homografie übergeben, die im nächsten Abschnitt aufgeführt wird. Im Rahmen dieser Studie soll zwischen quadratischen Produkten (Label: “square”), länglichen Produkten (Label: “long”) und Gläsern (“round”) unterschieden werden. Zudem soll die Hand (Label: “hand”) erkannt werden, um die Kundeninteraktion zu verfolgen. Aufgrund der überschaubaren Anzahl an Labels, sowie markante Unterschiede der Kategorie-Labels untereinander, soll das Modell auf einer Architektur trainiert werden, die vor allem schnelle Klassifizierungen durchführen kann.
Die Auswahl fiel auf das im Jahre 2020 publizierte YOLOv4-tiny von Jiang et al [9]. Basierend auf dem YOLOv4 Framework von Bochkovskiy et al. [8] nimmt YOLOv4-tiny Änderungen am Backbone vor, wodurch sich die Propagierung durch das Netz schneller gestaltet. Die YOLOv4-tiny erzielt auf dem MS COCO Datensatz eine mAP (mean average performance) von 38.1%, welches eine 8%ige Steigerung zum Vorläufermodell YOLOv3-tiny darstellt. Im Vergleich zu YOLOv4 läuft YOLOv4-tiny nahezu siebenmal schneller als YOLOv4, jedoch mit stärkeren Einbußen in der Genauigkeit (mAP). Da die Echtzeit-Objekterkennung auf moderater Hardware mit angemessenen Bilder pro Sekunde für diese Studie von Relevanz ist, wird YOLOv4-tiny für die Realisierung des Hybridmodells herangezogen.
Entgegen der vorgeschlagenen Datenaugmentierung des Redmon et al. Papers, wird in dieser Studie mit den Trainingsdaten lediglich eine horizontale Spiegelung, eine Helligkeitsvarianz um +/-10% und ein Zoomfaktor von bis zu 30% durchgeführt, womit der originale Datensatz von 248 Bildern auf 496 Bilder verdoppelt wird. Mit 257 Markierung ist das Label “hand” am stärksten im Datensatz repräsentiert, gefolgt 82 Label “long”, 80 Label “round” und mit 42 Label für “square”. Wie in den Ergebnissen zu sehen sein wird, sind die Ungenauigkeiten der Klassifizierung mit hoher Wahrscheinlichkeit auf die limitierten Trainingsdaten zurückzuführen.

Die Griffhöhe des Kunden soll anhand der Fläche der Bounding Box der Hand festgelegt werden. Da sich Hände in verschiedenen Zuständen befinden können (z.B. geschlossen oder

ausgestreckt), wird die Griffhöhe eines Kunden, im Gegensatz zur Originalstudie mit Tiefensensoren, nicht in kontinuierlichen Werten gemessen, sondern mit diskreten Werten. Um die kontinuierlichen Werte der gemessenen Handfläche auf eine Ordinalskala zu überführen (oberes Regal, mittleres Regal), wurden Schwellenwerte zur Zuordnung der Handfläche einer Regalhöhe definiert. Diese Schwellenwerte wurden anhand der Verteilung der Handgrößen zu den jeweiligen Regalhöhen bestimmt.

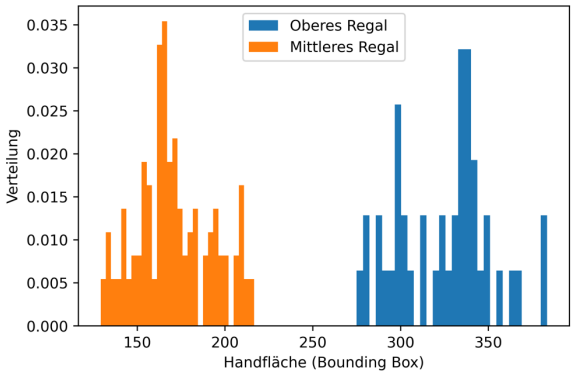

Abb.3: Zuordnung der Griffhöhe

Vor dem Regal wird eine Schranke markiert, die ein Durchschreiten von einer Hand als „hineingreifen“ interpretiert. Wird in das Regal mit leerer Hand gegriffen und mit einem Produkt verlassen, erkennt das Programm, dass ein Produkt aus dem Regal geholt wurde. Greift die Hand mit einem Produkt durch die Schranke und verlässt diese ohne Produkt, wird dies als „zurücklegen“ erkannt.

2.2.3 Klassifizierung der Objekte durch Feature Matching

Für die genaue Produkterkennung und Unterscheidung ähnlich aussehender Produkte wird ein Homografie-Verfahren mit dem Oriented FAST and rotated BRIEF (ORB) Algorithmus durchgeführt. Hierbei werden mit Faltungsoperationen Image Features erkannt. Image Features sind vom Algorithmus aufgegriffene Auffälligkeiten im Bild, welche sich aus Ankerpunkten und Deskriptoren zusammensetzen. Solche Deskriptoren enthalten Informationen zum erkannten Merkmal, während der Ankerpunkt die Position des aufgefundenen Merkmals beinhaltet [10].
Im Vergleich zu anderen Feature Detection Algorithmen wie SIFT, erzielt ORB eine vergleichbare Performanz zu einer deutlich geringeren Rechenzeit, weshalb sie für die vorliegende Echtzeit Anwendung gewählt wurde. [11]

Zur Unterscheidung von Produkten mittels ORB, werden die vom YOLO Modell erkannten Kategorien in den Bounding Boxes aus dem Frame extrahiert. Aus dem isolierten Abfragebild wird nun mit ORB die Feature Detection durchgeführt, um die markanten Stellen des gehaltenen Produkts ausfindig zu machen.

Im nächsten Schritt gilt es die Merkmale des Produkts in der Aufnahme auf Ähnlichkeiten mit einem Referenzbild (Trainingsbild) zu prüfen. Das sogenannte Feature Matching wird in dieser Studie mit dem Brute-Force Matcher durchgeführt. Dieser berechnet die Distanz eines Deskriptors des Abfragebildes zu allen Deskriptoren des Trainingbildes und gibt die Merkmale mit der niedrigsten Distanz zurück. Je niedriger die Distanz der Deskriptoren, desto stärker ähneln sich die Merkmale [12].
In diesem Versuch werden die Durchschnittsdistanzen der 20 besten Merkmale für die Produktklassifizierung verwendet. Der Brute-Force Matcher wird auf jedes Referenzbild eines Produkts innerhalb einer Kategorie angewandt, um die Ähnlichkeiten der Produkte zum Abfragebild zu ermitteln.

Das Produkt, dessen Referenzbild zum Abfragebild die größte Ähnlichkeit aufweist, wird als Label für das erkannte Objekt in der Aufnahme genommen.

3. Ergebnisse

3.1 Quantitative Auswertung

Um das Software Modell zu testen, wurde ein Testvideo mit derselben Kameraaufstellung und -einstellung wie bei den Trainingsbilder erstellt. In diesem Video werden unterschiedlich positionierte Produkte aus dem oberen und mittleren Regalfach genommen und wieder zurückgelegt. Es soll dabei abgelesen werden, wie zuverlässig die Produkterkennung funktioniert, als auch inwiefern die Kundeninteraktion durch Hand-Tracking abgebildet werden kann.

Zur Messung der Performanz der Produkterkennung, wird das Produkt angegeben, welches in einer Zeitspanne zu sehen ist. Über diesen Zeitraum wird gemessen, ob das richtige Produkt, das falsche Produkt in der richtigen Kategorie, das falsche Produkt in der falschen Kategorie, oder gar kein Produkt auf dem Frame von dem Modell erkannt wird. Aus diesen gewonnenen Daten, kann separiert betrachtet werden, wie akkurat die Objekterkennung des YOLOv4 Modells ist und wie sicher die Unterscheidung der Homografie funktioniert.

Das Ergebnis des Testdurchlaufs zeigt, dass auf 66,7% aller Testframes das richtige Produkt erkannt wurde. Auf 12,5% der Frames wurde die richtige Kategorie erkannt, jedoch wurde das Produkt durch die Homografie falsch klassifiziert. In 3,2% der Fälle wurde durch die YOLOv4 Objekterkennung die falsche Kategorie klassifiziert und in 17,2% der Frames wurde kein Objekt erkannt. Um die Performanz des YOLO-Modells zu ermitteln, werden alle richtig klassifizierten Kategorien betrachtet.

Insgesamt wurden 79,6% der Frames vom neuronalen Netz richtig erkannt. Für die Leistung der Homografie werden alle richtigen Kategorien, die in mehr als zwei Produkten untergliedert sind, mit allen richtig klassifizierten Produkten ins Verhältnis gesetzt. Die Homografie erzielt hierbei eine niedrigere Trefferquote von 66,7%.

Mit 66,9% und 66,4% ist das Verhältnis der richtig klassifizierten Produkte sowohl auf dem mittleren als auch oberen Regal ähnlich. Während die yolo Performance auf dem oberen Regal 88.7% erreicht, liegt diese bei dem mittleren Regalfach bei 74,3%. Auf der mittleren Regalhöhe ist das Verhältnis nicht-

erkannter Produkte mit 24,7% mehr als doppelt so hoch wie im oberen Regalfach (11,1%).

Auch ist das Verhältnis an falsch erkannten Kategorien auf mittlerer Regalhöhe (3,9%) doppelt so hoch wie beim oberen Regal (1,9%). Insgesamt nimmt die Fehlerquote in Bezug auf die Klassifizierung der Kategorie vom oberen zum mittleren Regalfach um 44,1% zu. Die Genauigkeit der Homografie auf Höhe des mittleren Regalfachs beträgt 78,9%, während sie beim oberen Regal lediglich bei 49,9% liegt.

	Richtiges Produkt	Richtige Kategorie aber falsches Produkt	Falsche Kategorie	Kein Produkt
Oberes Regal	66.98%	21.70%	1.89%	9.43%
Mittleres Regal	66.48%	7.82%	3.91%	21.79%
Gesamt	66.67%	12.98%	3.16%	17.19%

Abb.4: Testergebnisse

Bei genauerem Untersuchen des Testvideos wird sichtbar, dass vor allem abgeschnittene Produkte am Rand eines Frames oder Regals nicht aufgegriffen oder fehlklassifiziert werden. Wird die Messung nochmals durchgeführt, jedoch nur mit Frames in denen mindestens ¾ eines Produkts sichtbar ist, verbessern sich die Messeergebnisse erheblich. Die Klassifizierung richtiger Produkte steigt auf 80,0% an und das YOLO Modell erzielt eine Performanz von 86,9%. 3,7% der Produkte wurden weiterhin falsch klassifiziert, während auf 9,7% der Frames keine Produkte erkannt wurden.

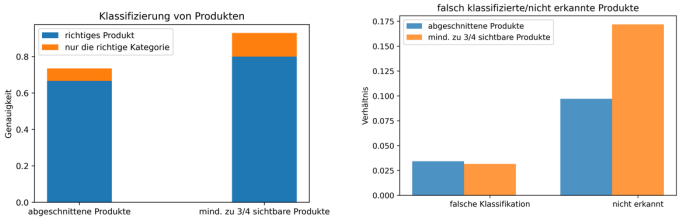

Abb.5: Klassifizierungsgenauigkeit aller Frames vs. Frames mit sichtbaren Produkten

Auf einem Intel i7 3770 lief die Objekterkennung im Durchschnitt mit 9.48 Bilder pro Sekunde.

3.2 Qualitative Auswertung

Trotz Schwächen in der Klassifizierung und Erkennung der Produkte, konnte die Interaktion des Kunden zuverlässig aufgezeichnet werden. Die Bestimmung der Höhe über die Größe des Begrenzungsrahmens der Hand und dem Schwellwert funktionierte im Testszenario. Die Ortung der Hand über die horizontale Achse verlief ebenfalls problemlos. Zum Testen wurde ein virtuelles Modell des Regals mit derselben Produktanordnung des Testvideos nachgebildet. Sobald eine Person die gesetzte Regalgrenze durchschreitet, werden die Stellen auf dem virtuellen Regal mit einem roten Punkt gekennzeichnet. Wie in der Originalstudie wird damit eine Heatmap der Kundeninteraktionen an einem Einkaufsregal nachgebildet [Abb.6]. Neben der Heatmap kann mit der Objekterkennung zusätzlich registriert werden, welche Produkte aus dem Regal genommen und wieder zurückgelegt werden. Hiermit konnten die drei Zustände der Originalstudie (positiv, negativ, neutral) rekonstruiert werden

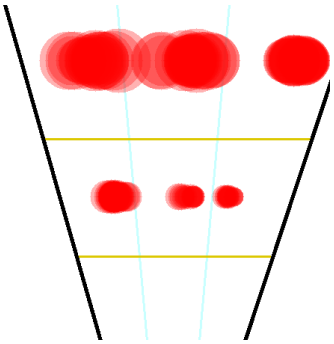

Abb.6: Heatmap

Mit der Objekterkennung muss die Kundeninteraktion nicht mehr durch eine initiale Produktanordnung zurückverfolgt werden, sondern kann direkt anhand der klassifizierten Produkte abgeleitet werden. Hiermit können beispielsweise auch fehlplatzierte Produkte im Regal erkannt werden.

Abb.7: Regalinteraktion

4. Diskussion / Ausblick

4.1 Ansätze zur Optimierung

Zur weiteren Optimierung des Modells kann die Genauigkeit durch eine Erhöhung der Menge an Trainingsdaten verbessert werden. Außerdem kann die Genauigkeit durch den Einsatz weitere Datenerweiterungstechniken erhöht werden. So können bisher verwendete Techniken wie Helligkeitsveränderungen verstärkt werden oder neue Techniken wie das Hinzufügen von Bildrauschen oder das Löschen zufälliger Bildpunkte hinzugefügt werden. Hierbei sollte Bedacht werden, dass Techniken wie Rotation des Bildes oder Veränderung des Hintergrundes die Genauigkeit in diesem Szenario kaum erhöhen werden, da die Kamera immer festmontiert ist und der Hintergrund immer derselbe sein wird. Da das Modell Schwierigkeiten aufwies, abgeschnittene Produkte zu identifizieren, sollte bei der Datenpräparation das Cutout als Data Augmentation Ansatz in Erwägung gezogen werden. Mit dieser Methode werden Bereiche eines

Trainingsbildes verdeckt, welche dem Neuronalen Netz beim Training dazu verleitet unvollständige Produkte richtig zu klassifizieren.

Des Weiteren können vortrainierte Gewichte verwendet werden, die durch bereits trainierte Modelle mit gleichen oder ähnlichen Trainingsdaten erstellt wurden. Die Ableitung der Griffhöhe über den Begrenzungsrahmen der Hand, hat für das Testszenario ausgereicht, jedoch ist dies lediglich ein naiver Ansatz und bedarf genauere Untersuchung. Die Perspektive im virtuellen Regal wurde mit Hilfe eines intuitiv gewählten Fluchtpunkts nachgestellt.

Im Bezug zum Szenario eines Einzelhandelsgeschäft können aus den gewonnen Daten viele verschieden Rückschlüsse gezogen, und zusätzliche Marketingaspekte analysiert werden. So könnte man zum Beispiel die durchschnittliche Dauer eines Kunden vor dem Regal, die Anzahl der Kunden in einem bestimmten Zeitintervall oder die Menge an Interaktionen mit Produkten am Regal gemessen und schließlich analysiert werden. Darüber hinaus können Defizite in der Produktplatzierung aufgezeigt, Rückschlüsse auf die Effektivität von Werbung für ein Produkt im Regal gezogen oder auch Cross-Selling Strategien identifiziert werden. All diese Aspekte können mit dem Einsatz dieser Methode und ohne zusätzliche Hardwareanschaffungen analysiert werden, um mehr Informationen bezüglich des Verhaltens von Kunden in Einzelhandelsumgebungen zu erhalten.

4.2 Grenzen der Umsetzung des Laborexperiments

Durch die ständige Optimierung des Laborexperiments wurde bereits ein realitätsnahes Umfeld des Regals geschaffen. Dennoch mussten Abstriche aufgrund von fehlenden Möglichkeiten und der Messbarkeit vorgenommen werden. Beispielsweise wurden nur drei Produkte betrachtet, welche mit 20cm Abstand voneinander im Regalfach platziert wurden. Dies entspricht nicht den gängigen Anordnungen im Einzelhandel, in welchen Produkte meist berührend nebeneinander aufgereiht werden. Zudem entsprechen einige Proportionen nicht den geläufigen Größenverhältnissen von Regalen. In Einkaufsmärkten befinden sich meist höhere Regale, ein Unterschied, welcher auch die Position der Kamera beeinflusst.

Genau wie die Höhe des gesamten Regals, sind auch die Abstände von Regalfächern je nach Produktgröße anders in Einzelhandelsgeschäften als in dem Laborexperiment. Meist trifft man auf wenig Leerraum zwischen den einzelnen Regalfächern. Allgemein lässt sich festhalten, dass die fehlende Tiefendimension bei der Umsetzung durch eine entsprechend höhere Bildqualität zu kompensieren ist.

Sowohl die Qualität der Aufnahme als auch die präzise Einstellung verschiedener Kameraparameter sind notwendig, um Ergebnisse zu erlangen, welche für weitere Umsetzungsschritte unerlässlich sind.
Das aufgestellte Experiment strebte zwar nach möglichst realitätsnahen Umständen, dennoch stellen sich Fragen in Bezug auf die Umsetzbarkeit für Einzelhändler.

Die Originalstudie erstellte ein Framework, welches die erhobenen Daten der RGB-D Kameras über kostengünstige Einplatinencomputer an einen Router sendet, um entsprechend über das Internet auf die Daten zuzugreifen.

Solch eine Echtzeitübertragung gestaltet sich umständlicher bei der Verallgemeinerung des Ansatzes durch den Einsatz herkömmlicher Kameras. Da die eingesetzten Videodaten aufgrund diverser Qualitätseinstellungen größer sind, würden zu starke Qualitätseinbußen über das Internet das Ergebnis erheblich verschlechtern. Außerdem stellt sich die Frage, ob der entstehende Umsetzungsaufwand vor allem im Hinblick auf die Erstellung von Trainingsdaten noch im Verhältnis zu dem erzeugten Mehrwert für Einzelhändler steht. Die Aufnahme der Trainingsdaten ist zeit- und kostenintensiv und muss stetig erneuert werden, da äußerliche Produktmerkmale meist von den Herstellern gewechselt werden.

Somit würden neue Schriftzüge, wie beispielsweise Gewinnspiele auf Verpackungen oder andere äußerliche Veränderungen zu weiterem Aufwand führen, da die Hersteller entsprechend die Neugestaltung den Einzelhändlern mitteilen müssten. Die neuen Bilddaten würden dann für die Homographie neu eingepflegt werden, sodass neue Merkmale bekannt sind.

Ferner besteht ein wesentlicher Unterschied zur Originalstudie in der Rechenintensität, da durch die Kombination der YOLO-Object Detection mit der Homographie ein erhöhten Rechenaufwand entsteht. Vor allem für die Homographie werden hochauflösende Bilddaten benötigt, wodurch Herausforderungen entstehen im Anbetracht auf den erhöhten Datendurchsatz, welcher im Einzelhandel vorliegen kann.

Aufgrund der Abstraktion durch Subtraktion der Tiefendimensionsdaten, sind im Umkehrschluss umfangreiche Bilddaten vonnöten, welche aufgrund deren Qualität ein höheres Datenvolumen aufweisen.

5. Literaturangaben

[1] Originalpaper: Liciotti, D. et al. (2015, 27. August). Shopper Analytics: a customer activity recognition system using a distributed RGB-D camera network. arXiv.org.

[2] Ives B, Cossick K, Adams D. (2019, 26. März). Amazon Go: Disrupting retail? SAGE Journals.

[3] Asifullah, K. et al. (2019, 17. Januar) A Survey of the Recent Architectures of Deep Convolutional Neural Networks. arXiv.org.

[4] Redmon, J. et al. (2018, 8. April). YOLOv3: An Incremental Improvement. arXiv.org.

[5] Redmon, J. et al. (2016, 25. Dezember) YOLO9000: Better, Faster, Stronger

[6] Jiao, L. et al. (2019, 11. Juli). A Survey of Deep Learning-based Object Detection. arXiv.org.

[7] Lin, T. et al.(2016, 9. Dezember). Feature Pyramid Networks for Object Detection. arXiv.org.

[8] Bochkovskiy, A. et al. (2020, 23. April). YOLOv4: Optimal Speed and Accuracy of Object Detection.

[9] Mao, H et al. (2016, August) Towards Real-Time Object Detection on Embedded Systems

[10] Rublee, E et al. (2011, November) ORB: an efficient alternative to SIFT or SURF

[11] Karami, E. et al. (2017, 7. Oktober). Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images. arXiv.org.

[12] OpenCV: Feature Matching. (2019, 9. Oktober). OpenCV.