DOI: 10.2195/lj_Proc_duemmel_de_201912_01
URN: urn:nbn:de:0009-14-49804

Effizientes Labeln von Artikeln für das Einlernen Künstlicher Neuronaler Netze

Efficient labelling of articles for teaching artificial neural networks

Johannes Dümmel*
Maximilian Hochstein*
Johannes Glöckle
Kai Furmans

Institut für Fördertechnik und Logistiksysteme (IFL)
Karlsruher Institut für Technologie (KIT)

Gefaltete Neuronale Netze (CNN) wurden in den letzten Jahren auf Grund ihrer hohen Erkennungsgenauigkeit sowie ihrer hohen Erkennungsgeschwindigkeit zunehmend in der Objekterkennung eingesetzt. Trotz einer schnellen und zuverlässigen Klassifizierung besteht einer ihrer größten Nachteile darin, dass das Training eines solchen Netzes sehr zeitaufwendig ist. Der Grund dafür ist, dass abhängig von der Komplexität des zu erkennenden Objekts mehrere hundert bereits klassifizierte Lernbilder benötigt werden. Das Erstellen dieser Lernbilder erfolgt bisher überwiegend manuell. Aus diesem Grund wurde am Institut für Fördertechnik und Logistiksysteme (IFL) ein Assistenzsystem entwickelt, welches das Einlernen neuer Objekte gegenüber der klassischen manuellen Methode um ein Vielfaches beschleunigt.

[Schlüsselwörter: Assistenzsystem, künstliche neuronale Netze, Objekterkennung, Tiefenbild, Label]

Convolutional neural networks (CNN) have been increasingly used in object detection in recent years due to their high detection accuracy and high detection speed. Despite fast and reliable classifications, one of the biggest disadvantages is that the training of such a network is very time consuming. The reason for this is that, depending on the complexity of the object to be detected, several hundred already classified learning images are required. Until now, the creation of these learning images was mainly done manually. For this reason, an assistance system was developed at the Institute for Material Handling and Logistics (IFL), which accelerates the learning of new objects considerably compared to the traditional manual method.

[Keywords: assistance system, artificial neural network, object detection, depth image, label]

1 Einleitung und Motivation

Der Umsatz des Versandhandels in Deutschland stieg zwischen 2009 und 2018 um 140,7 % von 30,0 Mrd. € auf 72,2 Mrd. € an [Sta18]. Das resultierende hohe Paketaufkommen muss durch die Logistikzentren der Händler effizient bearbeitet werden. Um die Wege innerhalb eines Warenlagers möglichst kurz zu halten, wird jedem Kommissionierer ein Bereich des Lagers zugeteilt. Da jeder Kommissionierer mehrere Aufträge gleichzeitig bearbeitet, werden die einzelnen Artikel nach dem Kommissioniervorgang durch einen Konsolidierer den eingegangenen Kundenaufträgen zugeordnet. Dieser Vorgang ist sehr zeitaufwendig und fehleranfällig. Deshalb automatisieren große Logistikunternehmen mit hohen Durchsätzen oftmals die Konsolidierung. Doch Schätzungen zufolge greifen 80% der Lagerhäuser auf manuelle Sortiersysteme zurück. Vor allem für kleine und mittelständische Betriebe lohnen sich die hohen Investitionskosten einer Automatisierung nicht [Fra18].

Für diese Betriebe wurde am Institut für Fördertechnik und Logistiksysteme (IFL) des Karlsruher Instituts für Technologie (KIT) der Konsolidierassistent entwickelt und experimentell nachgewiesen, dass dessen Einsatz Kommissionierzeiten um bis zu 30 % verringert. Grundlage des Systems bildet der Einsatz eines Gefalteten Neuronalen Netzes (CNN). Dabei analysiert das CNN die Bilder einer Kamera und erkennt die zu sortierenden Artikel. Der Konsolidierassistent ordnet daraufhin die Artikel den entsprechenden Kundenaufträgen zu. Bei der Bearbeitung eines Auftrags strahlt eine Projektionseinheit alle dazugehörenden Artikel an. Nutzer sind dadurch in der Lage, die angestrahlten Artikel schnell und mit geringem kognitivem Aufwand einem Kundenauftrag zuzuordnen und manuell in der entsprechenden Kiste abzulegen [Hoc17].

In der industriellen Nutzung von CNN werden anwendungsspezifische Objekte erkannt. Das Einlernen dieser Objekte ist bisher jedoch sehr zeitintensiv, da das Objekt auf jedem Lernbild markiert werden muss. Deshalb

DOI: 10.2195/lj_Proc_duemmel_de_201912_01
URN: urn:nbn:de:0009-14-49804

ist die industrielle Nutzung von CNN für die Objekterkennung noch nicht weit verbreitet. Zur Erkennung eines neuen Objekts benötigt das CNN mehrere hundert Lernbilder dieses Objekts. Bilder aus unterschiedlichen Perspektiven sowie vor unterschiedlichen Hintergründen erhöhen die Erkennungswahrscheinlichkeit aus unterschiedlichen Blickwinkeln und auf unterschiedlichen Oberflächen. Ein Lernbild besteht aus der Bilddatei selbst sowie aus einer zusätzlichen Datei. Diese enthält neben der Identifikationsnummer (ID) des zu lernenden Objekts dessen Koordinaten im Bild in Form einer rechteckigen „Bounding Box“. Diese Box inklusive der ID wird im Folgenden als Label bezeichnet. Die Erstellung eines Labels wurde am IFL durch die Entwicklung eines Assistenzsystems teilautomatisiert. Dabei liegt der Fokus auf der Reduktion der Aufnahmezeit durch eine Beschleunigung des Aufnahmevorgangs der Bilder sowie auf der Einführung einer automatisierten Labelmethode.

2 Stand der Technik

Seit dem Ende der 80er Jahre des zwanzigsten Jahrhunderts existiert ein anwendungsorientierter Forschungszweig mit dem Ziel, künstliche neuronale Netze (KNN) für industrielle Anwendungen nutzbar zu machen [Pod01]. Im Bereich der Bildverarbeitung werden vor allem CNN eingesetzt. Diese gefalteten neuronalen Netze gehören zu den rekurrenten bzw. feedback KNN, da ihr Output dem Netzwerk wieder zugeführt wird [Goo16]. Neben der Bildklassifizierung, welche ein Bild einer Klasse zuordnet [Kri12], ermöglichen CNN auch die Lokalisierung von Objekten in Bildern. Für die Erstellung der dafür benötigten Label existieren unterschiedliche Vorgehensweisen, auf welche im Folgenden näher eingegangen wird.

2.1 Manuelles Labeln

Auf dem Markt existieren mehrere Softwarelösungen, die das manuelle Labeln von Bildern vereinfachen. Dazu gehören beispielsweise LabelMe [Rus08], Colabeler [Col18], Dataturks [Dat18] oder RectLabel [Rec18]. Der Labelprozess beginnt mit dem Öffnen, bzw. dem Hochladen eines Bildes. Der Nutzer markiert daraufhin manuell das zu labelnde Objekt, indem es durch ein Rechteck möglichst enganliegend eingerahmt wird (siehe Abbildung 1), und weißt dem Objekt eine ID zu. Im nächsten Schritt wandelt die Software das visuelle Label in Koordinaten um und speichert diese zusammen mit der ID in einer separaten Datei. Dabei variieren, abhängig von dem verwendeten CNN, die Darstellungsform sowie das Dateiformat. Sorokin und Forsyth [Sor08] beschreiben einen Ansatz, um diese Aufgabe über Amazon Mechanical Turk an die Gemeinschaft der Online Arbeiter auszulagern. Jeder Arbeiter bekommt dabei einen festgelegten Betrag pro klassifiziertem Bild. Über Amazon Mechanical Turk angestellte Arbeiter von LabelMe benötigen im Schnitt 47 s, um ein Label auf einem Bild zu erstellen [Lab12]. Das im Konsolidierassistent verwendete CNN YOLOv2 [Red16] erfordert pro Artikel 200 Lernbilder (siehe Kapitel 5). Insgesamt entspricht das einem zeitlichen Aufwand von 156:40 min pro Objekt für das manuelle Labeln.

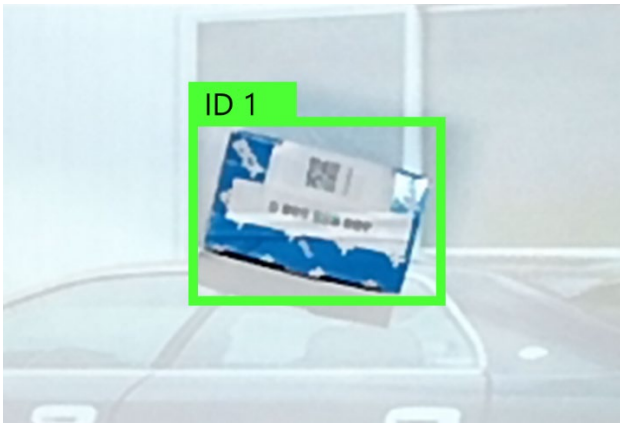

Abbildung 1: Originales Lernbild eines Artikels inklusive Hintergrundbild zur Steigerung der Varianz sowie Visualisierung des Labels inklusive ID.

2.2 Video Annotation

Liegen nicht nur einzelne Bilder eines Objekts vor, sondern ein Video, in welchem sich das Objekt bewegt, kann die Methode der Video Annotation eingesetzt werden. Dabei markiert der Nutzer das zu labelnde Objekt am Anfang einer gleichförmigen Bewegung sowie zum Zeitpunkt, an dem diese endet. Bewegt sich das Objekt im Video nicht gleichförmig, muss dessen Bewegung in kleinere, gleichförmige Abschnitte unterteilt werden. Der Nutzer labelt das Objekt dann am Anfang und am Ende jedes Abschnitts. Das Labeln erfolgt analog zum manuellen Labeln. Mit den Eingabedaten berechnet der Computer das Label der Bilder zwischen zwei Nutzermarkierungen automatisch. Im Gegensatz zum manuellen Labeln verringert dieser teilautomatische Ansatz den Arbeitsaufwand um die vom Computer berechneten Label. Video Annotation erfolgt beispielsweise mit der Software VATIC (Video Annotation Tool from Irvine, California) der Univesity of California [Von13], einer Weiterentwicklung von LabelMe.

2.3 Automatisierung durch die Nutzung von Tiefenbildinformationen

Pordel und Hellström [Por15] erreichen das teilautomatisierte Extrahieren polygoner Label von Objekten durch die Nutzung eines Kinect Kamerasystems. Die Tiefenbildkamera des Systems erkennt das ihr am nächsten gelegene Objekt. Daraufhin werden RGB- und Tiefenbild überlagert. Eine weiterentwickelte Version des Connected Component Labelling (CCL) [Sam88], das Extended Connected Component Labelling (ECCL) verbindet Pixel in dem so generierten Bild mit einer verbundenen Gruppe von Pixeln, wenn die Differenz der neuen Eintrittspixel zu den Pixeln der Gruppe geringer ist als ein definierter

DOI: 10.2195/lj_Proc_duemmel_de_201912_01
URN: urn:nbn:de:0009-14-49804

Schwellenwert. Sowohl vor dem Anwenden des ECCL als auch danach kann der Benutzer Anpassungen vornehmen, um die Ergebnisse des Algorithmus zu verbessern bzw. um geringfügige Korrekturen daran vorzunehmen. Der von Pordel und Hellström [Por15] vorgestellte Prozess erfordert trotz eines hohen Grades an Automatisierung manuelle Anpassungen der Label durch den Nutzer.

Eine weiterführende Automatisierung von Labelprozessen erreichen Lyubova und Filliat [Lyu12] durch die Nutzung von Tiefenbild- sowie RGB-Informationen einer Kinect durch einen humanoiden Roboter. Durch die Anwendung mehrerer Algorithmen werden Bildbereiche von Interesse vom Hintergrund separiert. Auf die dadurch erhaltenen Proto-Objects werden zwei weitere Algorithmen angewandt, die HSV3 Superpixel Segmentierung [Li16] sowie der Speeded-Up Robust Features (SURF) Algorithmus [Bay08]. Der Roboter erkennt damit Objekte in seinem Sichtfeld und weist diesen eine ID zu. Er unterscheidet damit auch seine eigenen von anderen Händen sowie von Objekten. Sobald ein Objekt einer ID zugewiesen ist, wird es bei wiederholtem Auftauchen innerhalb des Sichtfeldes erneut erkannt. Der von Lyubova und Filliat [Lyu12] vorgestellte Prozess labelt Objekte zuverlässig und ohne weitere Anpassungen durch den Nutzer. Aufgrund der Anwendung vieler unterschiedlicher Algorithmen ist die Implementierung desselben jedoch mit erheblichem Programmieraufwand verbunden. Außerdem erkennt das System Objekte lediglich auf homogenem Untergrund. Die Wahrscheinlichkeit der Wiedererkennung derselben Objekte auf anderen Oberflächen verringert sich damit erheblich.

3 Zielsetzung

Ziel des Projekts war die Entwicklung eines effizienten Prozesses für die Erstellung von Lernbildern für CNN. Dabei lag der Fokus einerseits auf der möglichst kurzen Bearbeitungszeit, andererseits auf der Vermeidung von Fehlern.

Die kurze Bearbeitungszeit wird durch das Umdrehen der Reihenfolge des Prozesses im Vergleich zum manuellen Labelprozess erreicht. Im hier vorgestellten Prozess wird zunächst das Label berechnet und danach das einzulernende Objekt fotografiert.

Die Vermeidung von Fehlern geschieht zum einen durch eine Teilautomatisierung des Labelprozesses und zum anderen durch das Anregen des Spieltriebes des Nutzers, der so genannten „Gamefication“. Gamefication wird definiert als die Verwendung von Spielelementen in einem nicht spielerischen Kontext [Det11], um Benutzer von Software zu einem bestimmten Verhalten bzw. zu bestimmten psychologischen Ergebnissen zu motivieren [Sti17] und eine monotone Arbeit abwechslungsreicher zu gestalten.

4 Funktionsweise des Systems

Der Konsolidierassistent (siehe Abbildung 2) wurde auf Basis des bereits abgeschlossenen Projektes „Packassistent“ entwickelt [Hoc16]. Beim „Packassistenten“ handelt es sich um ein Kooperationsprojekt des IFL mit den Firmen Bedrunka+Hirth - Gerätebau GmbH und Optimum - datamanagement solutions GmbH, welches über das Zentrale Innovationsprogramm Mittelstand (ZIM) finanziert wurde. Abbildung 2 zeigt den Aufbau des Konsolidierassistenten. Über einer höhenverstellbaren Arbeitsfläche (7) befindet sich eine Projektionseinheit (6) sowie das Kamerasystem Microsoft Kinect für Xbox One (8). Vor dem Nutzer auf der gegenüberliegenden Seite des Tisches ist ein Bilschirm angebracht (4). Die Projektionseinheit fungiert als zweiter Bildschirm und projiziert Informationen auf die weiße Tischoberfläche. Die Kommissionieraufträge (3) werden in den blauen und roten Kisten auf den Tisch gelegt (2) und dem richtigen Kundenauftrag (1) zugeordnet.

Abbildung 2: Dargestellt ist Aufbau des Konsolidierassistenten [Hoc17] zu Beginne eines Konsolidierprozesses. Die markierten Komponenten sind Konsolidierboxen (1), aktueller Auftrag in Kommissionierboxen (2), Puffer mit weiteren Kommissionieraufträgen (3), Bildschirm für Auftragsverwaltung (4), Ablage für Kommissionieraufträge (5), Projektionseinheit (6), interaktive Tischoberfläche (7), Tiefenbildkamera (8).

Das Erstellen der Lernbilder für die spätere Objekterkennung erfolgt in drei Schritten. Im ersten Schritt werden Bilder aufgenommen, in welchen das zu lernende Objekt auf der Tischoberfläche liegt (siehe Kapitel 4.1). Danach nimmt der Nutzer das Objekt in die Hand und bewegt es über der Tischoberfläche und sichtbar für die Kamera in unterschiedliche Richtungen, während das Objekt in der Hand rotiert wird (siehe Kapitel 4.2). Im letzten Schritt werden die auf dem Tisch aufgenommenen Bilder kontrolliert, da hier durch menschliches Versagen Fehler entstehen können (siehe Kapitel 4.3). Im Folgenden werden diese drei Schritte genauer erläutert.

DOI: 10.2195/lj_Proc_duemmel_de_201912_01
URN: urn:nbn:de:0009-14-49804

4.1 Stationäre Lernbilderstellung

Bei der stationären Lernbilderstellung werden Lernbilder auf der Tischoberfläche liegend erstellt. Diese Methode ist notwendig um eine möglichst große Varianz innerhalb der Lernbilder zu gewährleisten. Es hat sich herausgestellt, dass bei alleiniger Verwendung der dynamischen Lernbilderstellung (siehe Kapitel 4.2) die Erkennungsgüte während der Anwendung drastisch sinkt, sobald das Objekt nicht in der Hand gehalten wird. Aus diesem Grund wird ein bestimmter Anteil der Lernbilder mit Hilfe der stationären Methode erstellt. Charakteristisch für den Teilprozess ist, dass mittels eines Projektors das Label auf die Tischoberfläche projiziert und das Objekt in das Label gelegt wird (siehe Abbildung 3). Der stationäre Labelprozess findet im Vergleich zur manuellen Vorgehensweise in umgekehrter Reihenfolge statt:

1. Projektion des Labels auf die Tischoberfläche.
2. Positionierung des Objektes innerhalb des Labels.
3. Fotografie des Objektes und speichern der Positionsinformationen.

Begonnen wird der Prozess, indem der Nutzer die Länge und Breite des Objekts in eine Eingabemaske eingibt, damit die projizierten Label den Abmaßen des Objektes entsprechen (siehe Abbildung 3). Anschließend wird die Position des Labels innerhalb des Projektionsbereiches sowie dessen Ausrichtung zufällig festgelegt und auf der Tischoberfläche visualisiert. Da das eingesetzte CNN nur Label mit horizontal und vertikal zum Bild ausgerichteten Kanten verarbeiten kann, wird die ungefähre Orientierung des Objektes mit Hilfe einer Linie innerhalb des Labels dargestellt (siehe Abbildung 3). Die Erstellung des projizierten Labels erfolgt, indem das System auf Basis der Größeneingaben eine Mittellinie (siehe Abbildung 4) definiert. Anschließend wird die Mittellinie entsprechend der Zielausrichtung verdreht und ein vorläufiges Label um die Mittellinie platziert. Die Eckpunkte des vorläufigen Labels definieren im letzten Berechnungsschritt die Größe des finalen Labels.

Damit die Erkennungsgüte weiter gesteigert werden kann, wird die Varianz durch die Verwendung zusätzlicher Methoden erhöht. So wird neben der Änderung des Blickwinkels durch eine Änderung der Ausrichtung und Position des Objektes, in 30 % der Fälle während des Labelprozesses ein zufällig ausgewähltes Hintergrundbild auf die Tischoberfläche projiziert. Darüber hinaus findet eine zusätzliche Variierung der Lernbilder in weiteren 30 % der Fälle durch die Reduktion der Bildgröße statt.

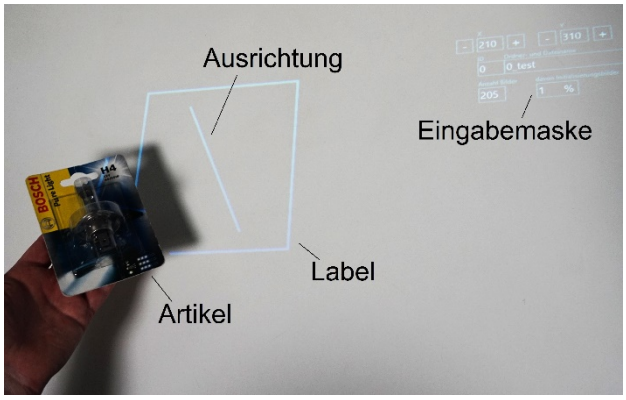

Abbildung 3: Dargestellt ist der Artikel mit dem dazugehörigen Label, die Mittellinie für die Ausrichtung sowie die Eingabemaske.

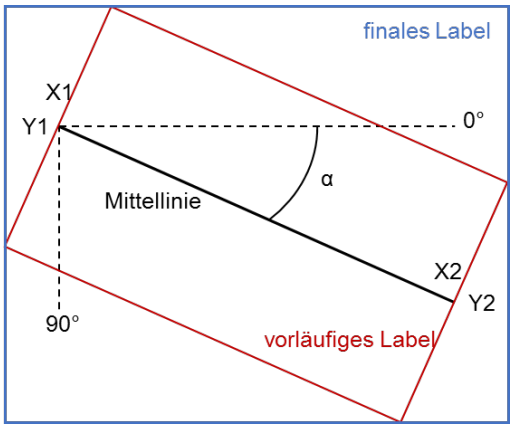

Abbildung 4: Skizziert ist das vorläufige (rot) sowie das finale Label (blau) inklusive der richtungsweisenden Mittellinie.

4.2 Dynamische Lernbilderstellung

Die dynamische Lernbilderstellung unterscheitet sich von der statischen Methode darin, dass die Lernbilder erstellt werden, während der Nutzer das Objekt in der Hand hält. Dies hat den Vorteil, dass die Lernbilder aus der Bewegung heraus und mit hoher Geschwindigkeit generiert werden können. Die Grundlage des Systems bildet der Einsatz einer Tiefenbildkamera zur Positions- und Formbestimmung des Objekts. Dafür kommt die am IFL entwickelte kommunikationsgestützte Lokalisierung zum Einsatz [Hoc19]. Das System analysiert das Tiefenbild und bestimmt den Schwerpunkt des Objekts, dessen Umriss, die Höhe oberhalb der Tischoberfläche, dessen Bewegungsrichtung als auch die Geschwindigkeit des Objekts (siehe Abbildung 5).

Während der Lernbilderstellung nimmt der Nutzer das zu lernende Objekt in die Hand und dreht es langsam unter der Kamera, um jede Seite des Objekts zu fotografieren. Dabei werden alle Objekte erkannt, die sich mindestens 15 cm über dem Tisch befinden. Für die weitere Auswertung dürfen sich nur die Hand des Nutzers sowie das darin befindliche Objekt in diesem Bereich befinden. Der minimale y-Wert *yMin* definiert die Oberkante des

DOI: 10.2195/lj_Proc_duemmel_de_201912_01
URN: urn:nbn:de:0009-14-49804

Labels. Im nächsten Schritt wird die Labeldiagonale aus der Nutzereingabe berechnet. Die tatsächliche Labeldiagonale hängt vom Abstand des Objekts zur Kamera ab. Die Höhe des Labels entspricht der Addition der tatsächlichen Labeldiagonale zu *yMin*. Daraufhin werden alle Begrenzungspunkte mit einem y-Wert größer *yMax* entfernt. Der minimale x-Wert der übrigen Begrenzungspunkte entspricht nun *xMin* und der maximale x-Wert entspricht *xMax*. Im letzten Schritt werden die Koordinaten in das für YOLOv2 verwendbare Format umgewandelt. Auch in diesem Schritt wird bei 30 % der Bilder ein zufälliges Hintergrundbild auf die Tischoberfläche projiziert sowie bei 30 % der Bilder deren Größe reduziert.

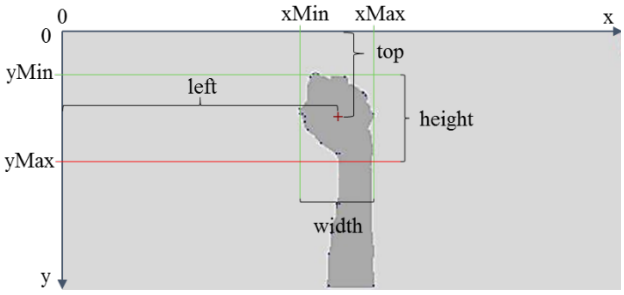

Abbildung 5: Topographisch dargestellt ist eine Hand inklusive Artikel, aufgezeichnet von der Tiefenbildkamera. Zu erkennen ist der Umriss der Hand mit den formbeschreibenden Begrenzungspunkten (blau) sowie das automatisch erstellte Label inklusive Labelmittelpunkt (rotes Kreuz).

Eine Herausforderung stellt dabei die Geschwindigkeit dar, mit der der Nutzer den Artikel bewegt. Eine zu hohe Geschwindigkeit resultiert in einer Bewegungsunschärfe, während eine zu geringe Geschwindigkeit zu einer geringen Varianz führt. Beides resultiert in einer minderen Qualität der Lernbilder, wodurch die Erkennungsgüte sinkt. Um das zu vermeiden wird die Geschwindigkeit permanent ermittelt und Lernbilder nur dann generiert, wenn die Geschwindigkeit in einem akzeptablen Bereich liegt. Durch den Einsatz von Gamification wird der Nutzer dazu animiert, das Objekt im idealen Geschwindigkeitsbereich zu bewegen. Abhängig von der Abweichung der aktuellen zur idealen Bewegungsgeschwindigkeit werden bei jedem Lernbild Punkte vergeben und aufsummiert. Der Nutzer wird dadurch angeregt möglichst viele Punkte zu sammeln, wodurch der Prozess kurzweiliger erscheint.

4.3 Labelkontrolle

Mit der Software „Labelkontrolle“ werden die zuvor statisch generierten Lernbilder (siehe Kapitel 4.1) auf Fehler überprüft. Dabei importiert die Software die Koordinaten des Labels in das dazugehörige Bild (siehe Abbildung 6). Auf diese Weise können Bilder aussortiert werden, auf welchen sich das Objekt nicht innerhalb des Labels befindet bzw. die Hand des Nutzers das Objekt verdeckt.

Abbildung 6: Bildschirmfoto während des Prozesses der Labelkontrolle. Zu sehen ist ein fehlerhaft gelabelter Artikel.

5 Validierung

Zur Validierung des gesamten Prozesses und der dazu gehörenden Software wurde das Labeln von Bildern in einem Laborexperiment untersucht. In Vorversuchen wurde die Anzahl der Lernbilder pro Objekt auf 200 festgelegt. Es stellte sich heraus, dass eine größere Anzahl an Bildern zu keiner signifikanten Verbesserung der Erkennungsgüte führt. Von den 200 Bildern sind 10 % stationäre (siehe Kapitel 4.1) und 90 % dynamische Lernbilder (siehe Kapitel 4.2).

24 Versuchspersonen durchliefen mit jeweils zwei Beispielartikeln den gesamten Labelprozess. Dieser besteht aus der Aufnahme und dem automatisierten Labeln der Artikel sowie der nachfolgenden Überprüfung der stationären Lernbilder mit Hilfe der Labelkontrolle. Es wurden pro Artikel 204 Bilder aufgenommen, 24 davon stationär und 180 dynamisch. Da unerfahrene Nutzer fehlerhafte Bilder produzieren, welche in der Labelkontrolle gelöscht wurden, erhöht sich die Anzahl der Tischbilder um einen Puffer von 4 Bildern.

Die Gesamtdauer des Prozesses verringerte sich bei 22 von 24 Probanden bei Artikel 2 (siehe Abbildung 7). Außerdem verringerte sich im Mittel die Gesamtdauer von 2:11 min bei Artikel 1 auf 1:56 min bei Artikel 2 (siehe Tabelle 1 und Tabelle 2). Der dabei vermutete Lerneffekt wurde in einem weiteren Versuch mit fünf Versuchspersonen und jeweils zehn Artikeln validiert.

DOI: 10.2195/lj_Proc_duemmel_de_201912_01
URN: urn:nbn:de:0009-14-49804

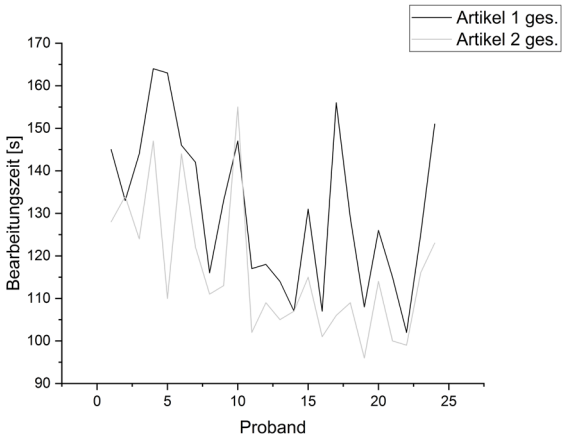

Abbildung 7: Bearbeitungszeit der Lernbilderstellung von 24 Probanden für jeweils zwei Artikel.

6 ERGEBNISSE UND AUSWERTUNG

Die Ergebnisse der Versuche von 24 Probanden mit jeweils zwei Artikeln werden im Folgenden diskutiert. Tabelle 1 zeigt die statistischen Kennwerte der Versuche mit Artikel 1 und Tabelle 2 zeigt die statistischen Kennwerte der Versuche mit Artikel 2. Die durchschnittliche Gesamtdauer des Prozesses sinkt von 2:11 min bei Artikel 1 auf 1:56 min bei Artikel 2. Außerdem verringert sich die durchschnittliche Anzahl an Fehlern und damit an gelöschten Bildern pro Artikel von durchschnittlich 2,38 auf 0,63. Die Maximale Anzahl an Fehlern pro Artikel verringert sich von 12 auf 3. Damit wird der Zielwert von mindestens 20 stationären Bildern langfristig durch den Lerneffekt erreicht.

Für die Validierung des Lerneffekts werden die Versuche der fünf Probanden nach deren Reihenfolge der Durchführung nummeriert. Der Korrelationskoeffizient des arithmetischen Mittelwerts der fünf Probanden beträgt -0,56. Folglich sinkt mit steigender Anzahl der Versuche die Dauer des Prozesses (siehe Abbildung 8). Aufgrund dieses Lerneffekts werden zum Vergleich des entwickelten Prozesses mit dem manuellen Labeln die Versuche mit Artikel 2 herangezogen. Resultierend verringert sich die Dauer des Gesamtprozesses im Vergleich zum manuellen Labeln von 156:40 min um 98,79 % auf 1:56 min.

Einschränkungen des entwickelten Prozesses bestehen bei der Größe und dem Gewicht des einzulernenden Objekts. Da beim Aufnehmen der Lernbilder das Objekt gehalten werden muss, sollten auf Grund ergonomischer Randbedingungen keine schweren und großen Objekte eingelernt werden.

Tabelle 1. Statistische Kennwerte der Lernbilderstellung für Artikel 1 bei 24 Probanden.

	Stat. [s]	Dyn. [s]	Kontr. [s]	Ges. [s]	Fehler
$\bar{x}$	63,21	31,38	36,21	130,79	2,38
σ	7,58	7,35	10,08	18,07	3,01
x_{min}	56	19	16	102	0
x_{max}	81	42	56	164	12

Tabelle 2. Statistische Kennwerte der Lernbilderstellung für Artikel 2 bei 24 Probanden.

	Stat. [s]	Dyn. [s]	Kontr. [s]	Ges. [s]	Fehler
$\bar{x}$	60,92	28,83	26,50	116,25	0,63
σ	5,38	7,57	6,97	15,44	0,86
x_{min}	55	19	17	96	0
x_{max}	76	55	42	155	3

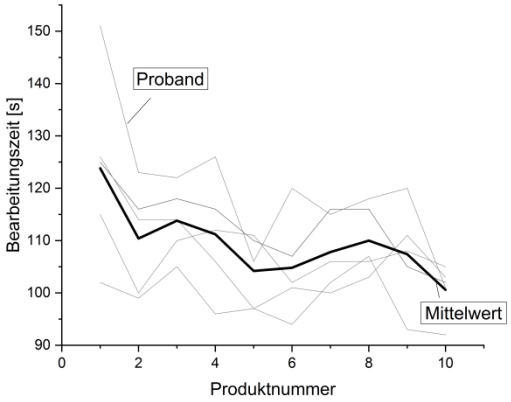

Abbildung 8: Gemessene Zeiten von fünf Probanden mit jeweils zehn Artikeln sowie deren arithmetischer Mittelwert zur Beurteilung der Lernkurve.

DOI: 10.2195/lj_Proc_duemmel_de_201912_01
URN: urn:nbn:de:0009-14-49804

7 Zusammenfassung und Ausblick

Durch den am IFL entwickelten Prozess reduziert sich die benötigte Zeit zum Labeln von Lernbildern für CNN aktuell um 98,79 % gegenüber dem manuellen Labeln. Die Versuche haben gezeigt, dass der entwickelte Prozess langfristig Lernbilder im Mittel in 1:56 min erstellt.

Eine Steigerung der Erkennungswahrscheinlichkeit kann durch die Separation von Vorder- und Hintergrund erreicht werden. Dafür kommen zwei Methoden in Frage. Durch das Übereinanderlegen von RGB- und Tiefenbild können alle Bildpunkte auf dem RGB-Bild entfernt werden, welche auf dem Tiefenbild nicht erscheinen, also unterhalb eines Grenzwertes liegen. Die zweite Möglichkeit zur Separation besteht in der Anwendung eines zum Zauberstab-Werkzeug in Adobe Photoshop ähnlichen Algorithmus. Dieser entfernt alle Pixel einer bestimmten Farbe oder Schattierung auf dem RGB-Bild. Hier kann beispielsweise der SIOX-Algorithmus eingesetzt werden [Fri05]. Um die Entfernung eines weißen Objekts auszuschließen, kann auch eine Kombination beider Methoden sinnvoll sein. Das damit verbundene Ausschneiden des Objekts hat den Vorteil, dass kein Hintergrundbild projiziert werden muss. Dieses fügt die Software vor dem Speichern jedem Bild hinzu. Falls es gelingt, lediglich das Objekt ohne Hand auszuschneiden, kann außerdem auf die Tischbilder verzichtet werden und es wird eine weitere Steigerung der Geschwindigkeit erreicht. Zusätzlich entfällt die Labelkontrolle. Damit müssen nur die Bilder mit dem Objekt in der Hand aufgenommen werden, was die Dauer signifikant verkürzt.

In weiterführenden Versuchen kann die Qualität der Lernbilder genauer bestimmt werden. Die Erkennungswahrscheinlichkeit wird durch das Einlernen des CNN mit den Lernbildern der Probanden ermittelt.

CNN werden in unterschiedlichen Bereichen zur Objekterkennung eingesetzt. Die Reduktion der Dauer des Labelprozesses steigert deren Nutzen, da das Erstellen von Lernbildern kostengünstiger und schneller erfolgt. Durch das Einlernen unterschiedlicher, anwendungsbezogener Objekte können CNN zukünftig erfolgreich in industriellen Anwendungen eingesetzt werden.

Literatur

[Bay08] Bay, H., A. Ess, T. Tuytelaars und L. van Gool (2008): SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding (CVIU) 110* (3), S. 346-359.

[Col19] Colabeler (2019): Colabeler. http://www.colabeler.com/. [abgerufen am: 28.06.2019].

[Dat19] Dataturks (2019): Dataturks. https://dataturks.com/. [abgerufen am: 28.06.2019].

[Det11] Deterding, S., D. Dixon, R. Khaled und L. Nacke (2011): From Game Design Elements to Gamefulness: Defining „Gamification“. In: *MindTrek '11 Proceedings of the 15th International*, S. 9-15.

[Fra18] Franzke, T. (2018): Der Mensch als Faktor in der manuellen Kommissionierung: Eine simulationsbasierte Analyse der Effzienz in Person-zur-Ware-Kommissioniersystemen. Wiesbaden: Springer Fachmedien Wiesbaden.

[Fri05] Friedland, G., K. Jantz und R. Rojas (2005): SIOX: Simple Interactive Object Extraction in Still Images. In: *Seventh IEEE International Symposium on Multimedia (ISM '05)*, S. 253-260. IEEE.

[Goo16] Goodfellow, I., Y. Bengio und A. Courville (2016): Deep learning. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press.

[Hoc16] Hochstein, M., J. Glöckle, T. Meyer und K. Furmans (2016): Packassistent - Assistenzsystem für die Qualitätskontrolle während des Packprozesses. Logistics journal, WGTL 2016.

[Hoc17] Hochstein, M., C. Kunert, J. Glöckle, M. Averweg, H. Weil und K. Furmans (2017): Konsolidierassistent - Assistenzsystem für manuelle Konsolidier- und Sortierprozesse in Distributionszentren. Logistics journal, WGTL 2017.

[Hoc19] Hochstein, M. (2019): Entwicklung und Evaluierung einer Kommunikationsgestützten Lokalisierung. Wird veröffentlicht von *KIT Scientific Publishing*

2019, (zugleich Dissertation Karlsruher Institut für Technologie 2019).

[Kri12] Krizhevsky, A., I. Sutskever und G. E. Hinton (2012): ImageNet Classification with Deep Convolutional Neural Networks. In: F. Pereira, C. J. C. Burges, L. Bottou, und K. Q. Weinberger (Hrsg.), *Advances in Neural Information Processing Systems 25*, S. 1097-1105. Curran Associates, Inc.

[Lab12] LabelMe (2012): Amazon Mechanical Turk for LabelMe. http://labelme.csail.mit.edu/Release3.0/browser-Tools/php/mechanical_turk.php. [abgerufen am: 28.06.2019].

[Li16] Li, D. und C. Liu (2016): Improved SLIC Superpixel Segmentation Based on HSV Nonuniform Quantization. In: *Proceedings of the 6th International Conference on Information Engineering for Mechanics and Materials*, Paris, France. Atlantis Press.

[Lyu12] Lyubova, N. und D. Filliat (2012): Developmental Approach for Interactive Object Discovery. In: *2012 International Joint Conference on Neural Networks (IJCNN 2012 - Brisbane)*, S. 1-7.

[Pod01] Poddig, T. und I. Sidorovitch (2001): Künstliche Neuronale Netze: Überblick, Einsatzmöglichkeiten und Anwendungsprobleme. In: H. Hippner, U. Küsters, M. Meyer, und K. Wilde (Hrsg.), *Handbuch Data Mining im Marketing*, Business computing, S. 363-402. Wiesbaden: Vieweg.

[Por15] Pordel, M. und T. Hellström (2015): Semi-Automatic Image Labelling Using Depth Information. *Computers 4* (2), S. 142-154.

[Rec19] RectLabel (2019): RectLabel. https://rectlabel.com. [abgerufen am: 28.06.2019].

[Red16] Redmon, J. und A. Farhadi (2016): YOLO9000: Better, Faster, Stronger. CoRR (abs/1612.08242).

[Rus08] Russell, B. C., A. Torralba, K. P. Murphy und W. T. Freeman (2008): LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision 77* (1-3), S. 157-173.

[Sam88] Samet, H. und M. Tamminen (1988). Efficient Component Labeling of Images of Arbitrary Dimension Represented by Linear Bintrees. *IEEE Transactions on Pattern Analysis and Machine Intelligence 10* (4), S. 579-586.

[Sta18] Statista (2019): Versandhandelsumsatz (inkl. E-Commerce) in Deutschland in den Jahren 2009 bis 2018 (in Milliarden Euro). https://de.statista.com/statistik/daten/studie/4452/umfrage/umsatzentwicklungim-versandhandel-seit-1980-in-deutschland/. [abgerufen am: 14.06.2019].

[Sor08] Sorokin, A. und D. Forsyth (2008): Utility data annotation with Amazon Mechanical Turk. *Proceedings of the first IEEE Workshop on Internet Vision at CVPR 2008*, S. 1-8.

[Sti17] Stieglitz, S., C. Lattemann, S. Robra-Bissantz, R. Zarnekow und T. Brockmann (Hrsg.) (2017): *Gamification: Using Game Elements in Serious Contexts*. Progress in IS. Cham: Springer International Publishing.

[Von13] Vondrick, C., D. Patterson und D. Ramanan (2013): Efficiently Scaling up Crowdsourced Video Annotation: A Set of Best Practices for High Quality, Economical Video Labeling. *International Journal of Computer Vision 101* (1), S. 184-204.

DOI: 10.2195/lj_Proc_duemmel_de_201912_01
URN: urn:nbn:de:0009-14-49804

M.Sc. Johannes Dümmel, Research Associate at the Institute for Material Handling and Logistics (IFL) at Karlsruhe Institute of Technology (KIT) in the department of Material Handling Technology. His research focuses on assistance systems and human-machine interaction.

Address: Institut für Fördertechnik und Logistiksysteme (IFL), Karlsruher Institut für Technologie (KIT),
Gotthard-Franz-Straße 8, 76131 Karlsruhe,
Tel.: +49 (0)721/608-48618,
E-Mail: johannes.duemmel@kit.edu

Dipl.-Ing. Maximilian Hochstein, Research Associate at the Institute for Material Handling and Logistics (IFL) at Karlsruhe Institute of Technology (KIT) in the department of Material Handling Technology. His research focuses on control technology and human-machine interaction.

Address: Institut für Fördertechnik und Logistiksysteme (IFL), Karlsruher Institut für Technologie (KIT),
Gotthard-Franz-Straße 8, 76131 Karlsruhe,
Tel.: +49 (0)721/608-48665,
E-Mail: maximilian.hochstein@kit.edu

M.Sc. Johannes Glöckle, software developer at SAP Deutschland SE & Co. KG. Johannes Glöckle studied computer science at the Karlsruhe Institute for Technology (KIT) from 2012 until 2018 and worked as a Research Assistant at the Institute for Material Handling and Logistics (IFL) from 2013 until 2015 and from 2016 until 2018. His focus lies in the field of image processing. Since 2018 he works for SAP Deutschland SE & Co. KG.

Address: SAP Deutschland SE & Co. KG, Hasso-Plattner-Ring 7, 69190 Walldorf,
E-Mail: johannes-gloeckle@t-online.de

Prof. Dr.-Ing. Kai Furmans, head of the Institute for Material Handling and Logistics (IFL) at Karlsruhe Institute of Technology (KIT).

Address: Institut für Fördertechnik und Logistiksysteme (IFL), Karlsruher Institut für Technologie (KIT),
Gotthard-Franz-Straße 8, 76131 Karlsruhe,
Tel.: +49 (0)721/608-48600,
E-Mail: kai.furmans@kit.edu

* Die gekennzeichneten Autoren haben zu gleichen Teilen zur Entstehung der Veröffentlichung beigetragen.