Artikel

Christian R. G. Dreher*, Manuel Zaremski, Fabian Leven, David Schneider, Alina Roitberg, Rainer Stiefelhagen, Michael Heizmann, Barbara Deml und Tamim Asfour

Erfassung und Interpretation menschlicher Handlungen für die Programmierung von Robotern in der Produktion

Capturing and Interpreting Human Actions for Programming Robots in the Production

Zusammenfassung: Der Mensch ist die flexibelste, aber auch eine teure Ressource in einem Produktionssystem. Im Kontext des Remanufacturings sind Roboter eine kostengünstige Alternative, jedoch ist deren Programmierung oft nicht rentabel. Das Programmieren durch Vormachen verspricht eine flexible und intuitive Alternative, die selbst von Laien durchführbar wäre, doch hierfür ist zunächst eine Erfassung und Interpretation von Handlungen des Menschen nötig. Diese Arbeit stellt eine multisensorielle, robotergestützte Plattform vor, welche die Erfassung zweihändiger Manipulationsaktionen, sowie menschlicher Posen, Hand- und Blickbewegungen während der Demontage ermöglicht. Im Rahmen einer Studie wurden an dieser Plattform Versuchspersonen bei der Demontage von Elektromotoren aufgezeichnet, um adäquate Datensätze für die Erkennung und Klassifikationen von menschlichen Aktionen zu erhalten.

Schlagwörter: Multisensorielle Erfassung des Menschen, Programmieren durch Vormachen, Blickrichtungsschätzung, Blickanalyse, Semantische Videorepräsentationen

***Korrespondenzautor: Christian R. G. Dreher, Tamim Asfour,** Karlsruher Institut für Technologie, Institut für Anthropomatik und Robotik, Hochperformante Humanoide Technologien, c.dreher@kit.edu, asfour@kit.edu

Manuel Zaremski, Barbara Deml, Karlsruher Institut für Technologie, Institut für Arbeitswissenschaft und Betriebsorganisation

Fabian Leven, Michael Heizmann, Karlsruher Institut für Technologie, Institut für Industrielle Informationstechnik

David Schneider, Alina Roitberg, Rainer Stiefelhagen, Karlsruher Institut für Technologie, Institut für Anthropomatik und Robotik, Maschinensehen für Mensch-Maschine Interaktion

Abstract: Human workers are the most flexible, but also an expensive resource in a production system. In the context of remanufacturing, robots are a cost-effective alternative, but their programming is often not profitable and time-consuming. Programming by demonstration promises a flexible and intuitive alternative that would be feasible even for non-experts, but this first requires capturing and interpreting the human actions. This work presents a multi-sensory robot-supported platform that enables capturing bimanual manipulation actions as well as human poses, hand and gaze movements during manual disassembly tasks. As part of a study, subjects were recorded on this platform during the disassembly of electric motors in order to obtain adequate datasets for the recognition and classification of human actions.

Keywords: Multi-Sensory Capturing of Human Actions, Programming by Demonstration, Gaze Estimation, Gaze Analysis, Semantic Video Representations

1 Einleitung

Der Mensch ist eine äußerst flexible, aber auch äußerst teure Ressource in einem Produktionssystem. Aus diesem Grund ist eine wettbewerbsfähige Produktion, die stark auf menschlicher Arbeit beruht, in Hochlohnländern nur für sehr werthaltige Produkte realisierbar. In dieser Arbeit betrachten wir den Anwendungsfall der Demontage von Elektromotoren. Dies steht im Kontext des Remanufacturing – d. h. der Demontage und Wiederaufbereitung von Gebrauchtprodukten. Dabei werden aktuell viele Prozesse manuell durchgeführt [35]. Herausforderungen bei der Automatisierung entstehen unter anderem durch ungewisse Produktzustände wie Beschädigungen, Rost, Verschmutzungen oder eine teilweise Zerlegung. Außerdem liegt meist eine hohe Variantenvielfalt vor und zudem fehlen aktuelle Konstruktionszeichnungen. Darüber hinaus ist eine maßgeschneiderte Roboterprogrammierung kostenintensiv und erfordert speziell ausgebildete Arbeitskräfte. In diesem Zusammenhang ist die Roboterprogrammierung durch Vormachen [5] ein vielversprechender Ansatz, um Laien zu ermöglichen, Roboter intuitiv durch Demonstrationen zu programmieren. Hierzu ist jedoch zunächst eine umfangreiche Erfassung des Menschen und eine Interpretationen der Demonstrationen notwendig. Insbesondere das Erkennen von Aktionen kann genutzt werden, um eine Demonstration in Aktionen und damit verbundenen Bewegungsprimitiven zu segmentieren. Diese wiederum können später auf einen Roboter reproduziert werden.

Wir stellen eine multisensorielle, robotergestützte Plattform vor, die eine Erfassung komplexer, zweihändiger Manipulationsaktionen, sowie menschlicher Posen, Hand- und Blickbewegungen während einer manuellen Demontage ermög-

licht (Abschnitt 3). Im Rahmen einer Nutzerstudie wurden an dieser Plattform Versuchspersonen bei der Demontage von Elektromotoren aufgezeichnet. Wir erläutern das Studiendesign und geben einen Überblick über die erhobenen Daten (Abschnitt 4). Anschließend stellen wir eine Methode zur zweihändigen Aktionserkennung vor, welche in die Station integriert werden soll (Abschnitt 5) und analysieren den aufgezeichneten Datensatz hinsichtlich der Erkennung von Handlungsabläufen (Abschnitt 6). Aufbauend auf den in der Nutzerstudie erhobenen Daten gehen wir anschließend auf die Externalisierung von implizitem menschlichen Handlungswissen mittels der Analyse von Augen- und Blickbewegungen (Abschnitt 7) ein.

2 Verwandte Arbeiten

2.1 Erfassung des Menschen

Hamabe et al. [17] präsentieren einen Aufbau zur Erfassung kollaborativer menschlicher Montage-Demonstrationen durch vier stationäre Kinect-Kameras. Aktionen des Menschen werden hierbei durch einen regelbasierten Ansatz erkannt. Parsa und Saadat [33] verwenden eine Plattform in einem kollaborativen Mensch-Maschine-Szenario bei der Demontage. Der Arbeitsbereich ist hierbei dreigeteilt, wobei ein Bereich von Mensch und Roboter geteilt wird, und die zwei verbleibenden Bereiche jeweils exklusiv dem Menschen oder dem Roboter zur Verfügung stehen. Anhand eines CAD-Modell des zu demontierenden Produkts wird ein Demontagevorranggraph erstellt, durch den das Sequenzplanungsproblem adressiert werden kann. Rakita et al. [36] lösen ein Optimierungsproblem zur Regelung eines Roboterarms mit montierter Kamera. Die Arbeit ist in der Telemanipulation angesiedelt, wobei ein Nutzer einen weiteren Roboterarm für Manipulationsaufgaben steuert. Der Roboterarm mit montierter Kamera wird hierbei so ausgerichtet, dass Verdeckungen mit dem vom Nutzer gesteuerten Roboterarm vermieden werden. Weiterhin findet in der Literatur das sogenannte Visual Servoing Einsatz, wobei ein Roboterarm mit montierter Kamera im Bildraum nachgeregelt wird [7, 32, 31]. In dieser Arbeit lösen wir das Problem der Ausrichtung des Roboterarms durch die Lokalisierung der menschlichen Hände im Arbeitsraum mittels einer zusätzlichen, stationären Kamera. Am Endeffektor des Robotearms ist eine zweite Kamera befestigt, um eine Nahaufnahme der Demontage zu erhalten.

Außerdem stellen wir in dieser Arbeit einen Ansatz für einen Aufbau zur Schätzung der Blickrichtung im Kontext einer manuellen Demontage ohne zutragende Sensorik mittels Video-Okulografie vor (siehe Unterabschnitt 3.3). (Der

Ansatz löst einige Herausforderungen, ist jedoch noch nicht zu Ende entwickelt, weshalb für Teile dieser Arbeit auch auf ein kommerzielles System mit kopfgetragener Sensorik zurückgegriffen wird.) Es gibt zahlreiche Methoden zur Schätzung der Blickrichtung ohne zutragende Sensorik mittels Video-Okulografie, siehe bspw. Hansen et al. [18]. Zhang et al. [56] stellen einen Vergleich verschiedener Methoden an und die Ergebnisse legen nahe, dass Blickregistrierungssysteme, die eine Beleuchtung im nahen Infrarot (NIR) verwenden und auf Reflexionen dieser Beleuchtung an der Cornea setzen, in unserem betrachteten Anwendungsfall genauere Schätzungen liefern könnten als solche, die dies nicht tun. Typische Szenarien, in denen entsprechende Aufbauten betrieben werden, sind allerdings am Bildschirm oder beim Führen eines Kraftfahrzeugs, siehe bspw. Holmqvist et al. [21], bei denen sich die Anforderungen von denen bei einer manuellen Demontage unterscheiden (für eine genauere Beschreibung der anwendungsspezifischen Anforderungen, bspw. in Bezug auf die Platzierung der Sensorik, um Verdeckungen durch Augenlider bei steilen Blicken zu minimieren, siehe Unterabschnitt 3.3). Siegfried et al. [42] stellen eine robotergestützte Plattform vor, die Ähnlichkeiten zu unserer aufweist. Es ist ein Blickregistrierungssystem, das ohne zu tragende Sensorik auskommt, integriert. Dieses setzt auf eine handelsübliche RGB-D-Kamera (anstatt beispielsweise auf eine maßgeschneiderte Kombination aus Kamera und NIR-Beleuchtung). Die Autoren schlussfolgern, dass weitere Forschung nötig sei, um bei einem Dutzend Objekten auf einem Tisch unterscheiden zu können, welches betrachtet wird.

2.2 Aktionserkennung

Bezüglich der Aktionserkennung existieren nebst regelbasierten Ansätzen [17] auch einige komplexe Arbeiten im Kontext von Programmieren durch Vormachen. Aksoy et al. [1] stellen ein Verfahren vor, wobei Aktionen anhand von Änderungen räumlicher Relationen zwischen semantischen Bildsegmenten beschrieben und modelliert werden. Das vorgestellte Verfahren ist in der Lage, eine so gebildete Aktionsrepräsentation auch zur Erkennung zu nutzen. Weitergeführt wurde diese Arbeit durch Ziaeetabar et al. [57], vor allem dadurch, dass in dieser Arbeit die räumlichen Relationen zwischen 3D-Objektmodellen berechnet wurden, statt zwischen 2D-Bildsegmenten. Koppula und Saxena [28] stellen ein Verfahren vor, bei dem die menschliche Pose und in der Szene vorkommende Objekte genutzt werden, um das Problem der Aktionserkennung durch ein Conditional Random Field zu adressieren. Auch in unserer Arbeit benutzen wir, ähnlich zu verwandten Arbeiten, räumliche Relationen, um eine Szene zu repräsentieren (bspw. [57], mit räumlichen Relationen ursprünglich vorgestellt in [2]). Diese Repräsentation

dient als Eingabe für einen Graphnetzwerk-Klassifizierer. Im Unterschied zu verwandten Arbeiten wird hierbei besonderer Fokus auf die Fähigkeit gesetzt, zweihändige Aktionen erkennen zu können. Weiterhin wird in dieser Arbeit die Pose des Menschen als neues Merkmal für eine bestehende zweihändige Aktionserkennung aus einer früheren Arbeit [9] verwendet.

In Abschnitt 6 führen wir eine qualitative Analyse von Clustern von Videosequenzrepräsentationen durch. Die Extraktion deskriptiver Repräsentationen zur Erkennung von Aktionen wie in dem von uns genutzten Ansatz von Schneider et al. [40] ist Bestandteil einer Vielzahl von Arbeiten, zum Beispiel mittels Vortraining auf reinen Videodaten [53, 24] oder kombiniert auf Videodaten gepaart mit Optischem Fluss [46]. Ein Beispiel für die Interpretation von Aktivitäten als Sequenzen von Repräsentationvektoren ist die Arbeit von Jamal et al. [23]. Im Gegensatz zu unserer qualitativen Analyse nutzen sie die Ähnlichkeit von Trajektorien in unterschiedlichen Repräsentationsräumen zur Berechnung einer Transformationsmatrix zwischen verschiedenen Domänen mittels maschineller Lernverfahren.

3 Erfassung des Menschen

Um die für die Reproduktion notwendigen Informationen aus menschlicher Demonstration abzuleiten, ist eine multisensorielle Erfassung des Menschen notwendig. Einerseits wird eine Kamera benötigt, um ein allgemeines Gesamtbild des Menschen sowie seines Arbeitsbereichs zu erhalten. Eine weitere Kamera wird benötigt, um eine detaillierte Nahaufname des Demontagevorgangs sicherzustellen. Diese Kamera wurde an den Endeffektor eines Franka Emika Panda Roboterarms montiert, welcher während einer Demontagedemonstration kontinuierlich so ausgerichtet wird, sodass den Handbewegungen des Menschen gefolgt wird. Dabei wird durch einen Mindestabstand sichergestellt, dass der Mensch während seiner Arbeit nicht physisch beeinträchtigt wird. Aus beiden Perspektiven sind weiterhin Tiefeninformationen von besonderer Bedeutung, da die Pose des Menschen im Arbeitsraum sowie die räumlichen Beziehungen zwischen relevanten Objekten, Werkzeugen, sowie den menschlichen Händen für die später folgende zweihändige Aktionserkennung als Eingabe benutzt werden (Abschnitt 5). Die Positionen der Hände werden weiterhin benutzt, um den Roboterarm mit der beweglichen Kamera auszurichten. Um später das Lernproblem für die zweihändige Aktionserkennung zu vereinfachen, und wichtige Zeitpunkte während der Demonstration zu erkennen, wird zusätzlich auf eine stationäre Blickregistrierung gesetzt. Zur Datenerfassung, -verarbeitung und Steuerung des Roboters wird das

Roboterentwicklungsframework ArmarX [47] verwendet. In Abbildung 1 wird der soeben beschriebene Versuchsaufbau veranschaulicht.

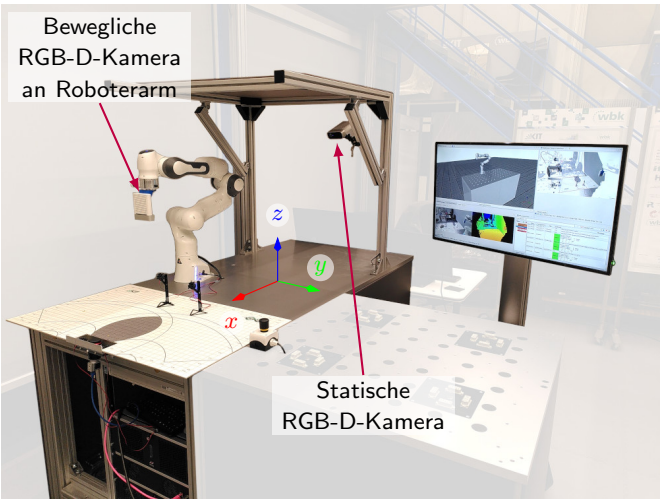


Abb. 1: Versuchsaufbau im Kontext des Projekts AgiProbot mit zwei RGB-D-Kameras. Eine Kamera ist statisch, die andere ist am Endeffektor eines Roboterarms befestigt. Mittig des Tisches ist die Koordinatensystemausrichtung des Arbeitsraums dargestellt. Der Tisch ist am Koordinatensystem ausgerichtet.

3.1 Erfassung der menschlichen Pose im Arbeitsraum

Zur Erfassung des Menschen im Arbeitsraum wurde die Integration von OpenPose [6] in ArmarX benutzt. Hierbei werden zuerst die Posen-Schlüsselpunkte des Menschen im Bildraum der statischen RGB-D-Kamera ermittelt, indem OpenPose auf dem RGB-Bild angewandt wird (vgl. Abbildung 2, ②). Die Ausgabe nach diesem Schritt ist die Pose des Menschen im Bildraum, dargestellt als Menge von annotierten Schlüsselpunkten (x, y, a), wobei (x, y) die Koordinaten des Schlüsselpunkts im Bildraum darstellt, und a eine Annotation des Punkts um das Gelenk zu beschreiben (bspw. „Linke Schulter", „Rechtes Handgelenk", „Nacken", ...). Mithilfe des Tiefenkanals der RGB-D-Kamera kann die 3D-Pose des Menschen im Koordinatensystem der Kamera ermittelt werden. Als letzter Schritt erfolgt eine Transformation der menschlichen Pose in den Arbeitsraum (vgl. Abbildung 2, ③).

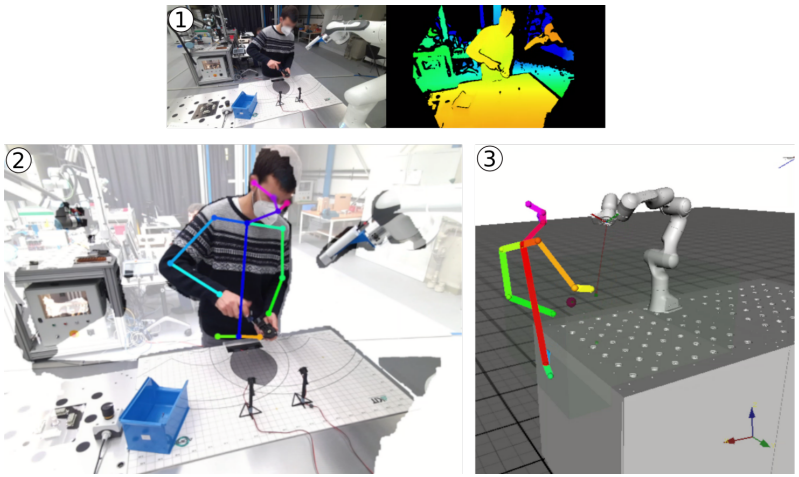


Abb. 2: Übersicht der Rohdaten und abgeleitete Daten der stationären RGB-D Kamera. ① RGB-D-Rohbild der statischen RGB-D-Kamera. ② Erkannte menschliche 2D-Pose von OpenPose angewandt auf das RGB-Bild. ③ 3D-Rekonstruktion der menschlichen Pose im Arbeitsraum aus erkannter 2D-Pose und Tiefenbild.

3.2 Erfassung der menschlichen Handbewegungen, Objekte und Werkzeuge

Zur Erfassung von Details während der Demontage ist eine möglichst nahe Erfassung des Demontagevorgangs notwendig. Hierfür wurde eine Microsoft Azure Kinect RGB-D Kamera an den Endeffektor eines Franka Emika Panda Roboterarms montiert. Das Koordinatensystem des Arbeitsraums ist in Abbildung 1 abgebildet.

Die Zielpose des Endeffektors $(p_{\text{target}}, R_{\text{target}})$ im Arbeitsraum ergibt sich aus der Zielposition p_{target} und der Zielorientierung des Endeffektors R_{target}. Der Berechnung der Zielposition p_{target} liegen zwei Parameter zugrunde: 1) der Mittelpunkt $\bar{h} = \frac{h_{\text{r}}+h_{\text{l}}}{2}$ beider Handpositionen h_{r} und h_{l}, sowie 2) ein Versatzvektor o des Endeffektors zum Mittelpunkt der Handpositionen $\bar{h}$. In unserem Versuchsaufbau wurde der Versatzvektor o so gewählt, dass er im Arbeitsraum an der x-Achse um 30° gedreht ist, an der y-Achse um −30° gedreht ist, und dass ein Abstand zum Mittelpunkt der Handpositionen $\bar{h}$ von 40 cm besteht. Dies schließt einerseits einen Sicherheitsabstand zum Menschen ein, und andererseits auch den Mindestabstand der Azure Kinect von 30 cm, da der Sensor für niedrigere Distanzen keine Tiefendaten liefert. Die Zielposition des Endeffektors p_{target} berechnet sich dann aus $p_{\text{target}} = \bar{h} + o$. Dies ist dargestellt in Abbildung 3. Die Zielorientierung des Endeffektors R_{target} ist viel schneller

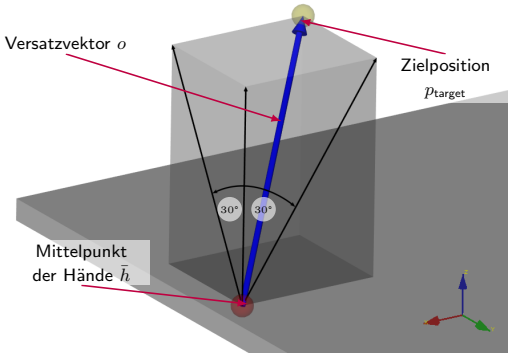


Abb. 3: Veranschaulichung des Versatzvektors o (blau) als Abstand des Roboter-Endeffektors zu den Händen des Menschen. In unserem Versuchsaufbau ist der Versatzvektor so gewählt, dass er $40\,\text{cm}$ lang ist, sowie an der x-Achse um $30°$, und an der y-Achse um $-30°$ gedreht ist.

einnehmbar als die Zielposition p_{target}, da weniger Gelenke benötigt werden und keine Strecke im Arbeitsraum zurückgelegt werden muss. Um bei schnellen Bewegungen die Hände des Menschen möglichst im Bild zu halten, wird aus diesem Grund die Zielorientierung zusätzlich so vorgegeben, dass die Kamera stets auf $\bar{h}$ ausgerichtet ist, während unter Umständen die Zielposition p_{target} erst noch eingenommen wird. Die Ausrichtung der Kamera wird ermittelt durch $\bar{h} - p_{\text{current}}$, wobei p_{current} die aktuelle Position des Endeffektors ist. Weiterhin wird der Rollen-Freiheitsgrad des Endeffektors so gewählt, dass die Kamera stets horizontal ausgerichtet bleibt. Die Zielpose $(p_{\text{target}}, R_{\text{target}})$ wird kontinuierlich basierend auf neuen Handpositionsdaten mit ca. $10\,\text{Hz}$ aktualisiert. Der Roboter wird schließlich mithilfe eines kartesischen Reglers in ArmarX auf die Zielpose $(p_{\text{target}}, R_{\text{target}})$ geregelt.

Auf den Bildern der beweglichen RGB-D-Kamera wird die Objekterkennung mit Hilfe von Detectron2, eine Bibliothek zur Objektdetektion und -segmentierung [52], durchgeführt, um Werkzeuge oder Einzelteile der Objekte erkennen zu können. Mittels Tiefeninformationen der beweglichen RGB-D-Kamera und der Vorwärtskinematik des Roboters können die Objekte im Arbeitsraum lokalisiert werden. Die 3D-Pose des Menschen, sowie erkannte Objekte und Werkzeuge dienen später als Eingabe zur zweihändigen Aktionserkennung (Abschnitt 5). Die hier beschriebenen verarbeitenden Komponenten zur Steuerung des Roboters, sowie Detectron2 zur Erkennung von Einzelteilen, wurden in ArmarX integriert.

Der vorgestellte Versuchsaufbau erwies sich als robust zum Verfolgen der Hände. Sporadische Sprünge der erkannten Handgelenkspositionen und dem dar-

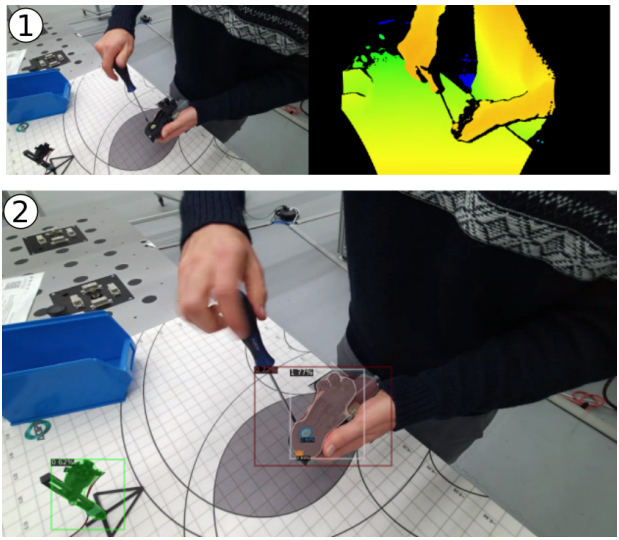


Abb. 4: Übersicht der Rohdaten und abgeleitete Daten der beweglichen RGB-D Kamera. ① RGB-D-Rohbild der beweglichen RGB-D-Kamera. ② Objekterkennung (Detectron2) angewandt auf ein RGB-Bild der beweglichen RGB-D-Kamera zur Erkennung von Werkzeugen und Einzelteilen.

aus berechneten Mittelpunkt werden durch den kartesischen Regler ausgeglichen. Die Parametrisierung des Versatzvektors ist so gewählt, dass eine zuverlässige Erfassung der zu erwartenden Aktionen erfolgen kann. Bewegungen, welche eine große Spannweite der Arme in Richtung des Versatzvektors einschließen, könnten problematisch sein, kommen praktisch jedoch nicht vor (siehe Abschnitt 4).

3.3 Erfassung der Blickrichtung

Zusätzlich zur menschlichen Pose und den menschlichen Handbewegungen wird die menschliche Blickrichtung mittels Video-Okulografie erfasst. Hierzu kann entweder ein kopfgetragenes (mobiles) oder fest-montiertes (stationäres) Blickregistrierungssystem zum Einsatz kommen. Im Rahmen der Forschungsstudie, die in Abschnitt 4 vorgestellt wird, wurde ein kommerzielles kopfgetragenes Blickregistrierungssystem verwendet. Die Erkenntnisse aus der Studie dienen neben der Analyse von implizitem menschlichem Handlungswissen auch als Anhaltspunkt zur Entwicklung eines anwendungsspezifischen, stationären Blickregistrierungssystems, welches es ermöglichen soll, dass keine Sensorik mehr vom Menschen getragen werden muss. Der Aufbau eines stationären Blickregistrierungssystems

im Kontext einer manuellen Demontage ist jedoch herausfordernd und nicht umfassend erforscht, wie auch in Abschnitt 2 näher erläutert wird.

Wir verwenden einen Aufbau mit einer Beleuchtung im nahen Infrarot (NIR). NIR-Licht hat den Vorteil, dass es nicht blendet und dass typische Kamerasensoren es detektieren können. Darüber hinaus sind folgende wichtige Aspekte zur NIR-Beleuchtung zu nennen. Erstens führt eine gute Ausleuchtung zu einer besseren Aufnahmequalität. Zweitens ermöglicht eine gute Ausleuchtung die Verwendung einer kleineren Blendenöffnung und führt zu einer erhöhten Schärfentiefe. Dies ist bei einer manuellen Demontage von besonderer Bedeutung, da eine verhältnismäßig große Bewegungsfreiheit besteht (im Vergleich zu bspw. der Arbeit an einem Bildschirm) und sich dadurch die Augen bei zu geringer Schärfentiefe aufgrund von Kopfbewegungen aus dem Schärfebereich der Kamera bewegen könnten. Drittens erzeugen NIR-Lichtquellen Reflexionen an der menschlichen Cornea-Oberfläche (Cornea-Reflexionen), welche direkt zur Schätzung der Blickrichtung verwendet werden können.

Es ist bekannt, dass zur Schätzung der Blickrichtung mittels Cornea-Reflexionen basierend auf geometrischen Überlegungen eine Kamera ausreichend ist (im Kontrast zu bspw. einem Stereo-Kamerasystem) und mindestens zwei Cornea-Reflexionen für die Kamera sichtbar sein müssen [16, 51]. Zusätzlich muss die Pupille für die Kamera sichtbar sein.

Eine besondere Herausforderung stellt in diesem Zusammenhang auch die mögliche Verdeckung der Sicht zwischen Auge und Kamera oder NIR-Lichtquellen dar. Diese kann durch Hände, Arme, Werkzeug oder andere Gegenstände unterbrochen sein. Aber auch Augenlider können bei zu steilen oder zu flachen Blicken für Verdeckungen sorgen. Damit Cornea-Reflexionen für die Kamera sichtbar sind, muss zusätzlich die Reflexionsbedingung für den Strahlengang von einer NIR-Lichtquelle zur Kamera an der Cornea-Oberfläche erfüllt sein, was weitere geometrische Nebenbedingungen an die Platzierung der NIR-Lichtquellen und der Kamera stellt. Darüber hinaus können die Kamera sowie die NIR-Lichtquellen nicht im Arbeitsvolumen platziert werden, da sie sonst bei der manuellen Demontage stören würden.

Abbildung 5 zeigt einen Aufbau, der den erwähnten Anforderungen während einer manuellen Demontage genügt. Zu sehen ist eine Arbeitsmatte mit einem eingezeichneten zweidimensionalen Schnitt eines Greifraums. Die Anbringung der Sensorik und NIR-Lichtquellen orientiert sich an den eingezeichneten Bereichen. Der dunkelgrau-schattierte Bereich zeigt den bevorzugten Bereich für zweihändiges Greifen. Die Kamera ist zwischen diesem Bereich und der Hüfte des beobachteten Menschen platziert. Es sind vier NIR-Lichtquellen angebracht – zwei davon neben der Kamera und zwei außerhalb des Greifraums. Die Szene zeigt das Lösen einer Mutter von einem Motor. Fünf Referenzpunkte (A - E) sind

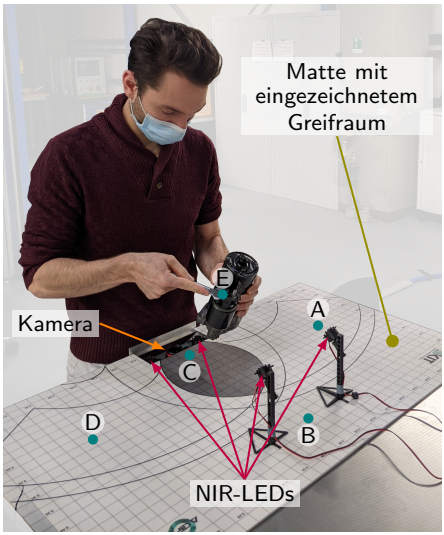


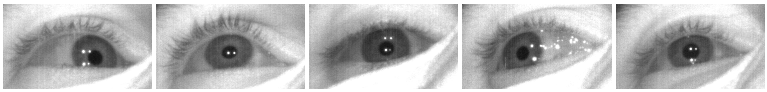
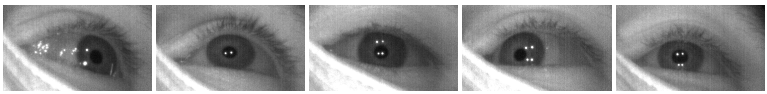
Abb. 5: Überblick zum Aufbau zur stationären Blickrichtungsschätzung während einer manuellen Demontage. Die Referenzpunkte A - E dienen als Blickziele zur Evaluation des Aufbaus.

als Blickziele markiert. Die Blickziele sind exemplarisch so gewählt, dass eine gewisse Abdeckung des Arbeitsvolumens zur Evaluation des Aufbaus erreicht wird. Die Blickziele A - D befinden sich auf der Tischoberfläche und verteilen sich auf den eingezeichneten zweidimensionalen Schnitt eines Greifraums. Das Blickziel E liegt auf der zu lösenden Mutter. In Tabelle 1 sind Bildausschnitte des linken und rechten Auges gezeigt, während diese auf die Blickziele fixiert waren. Kopfbewegungen wurden hierbei nicht restringiert. Man erkennt, dass in jedem Fall die Pupille sowie mindestens zwei Cornea-Reflexionen sichtbar sind. Somit ist eine stationäre Blickrichtungsschätzung für Blickpunkte innerhalb des durch die Blickziele A - E begrenzten Volumens mit dem gezeigten Aufbau prinzipiell möglich.

4 Empirische Forschungsstudie

Zur Untersuchung der beschriebenen Zielstellung und zur Erhebung eines umfangreichen Datensatzes wurden an der multisensoriellen, robotergestützen Plattform spezifische Demontageprozesse empirisch erforscht. Dazu erfolgte die Planung eines zielführenden Studiendesigns. Insgesamt haben an der Studie 15 Versuchs-

Tab. 1: Das Gesicht wurde in der Szene gemäß Abbildung 5 mit der markierten Kamera nahe der Hüfte aufgenommen. Die Tabelle zeigt Bildausschnitte des linken und rechten Auges während die Blickziele A - E fixiert werden.

Blickziel	A	B	C	D	E
rechtes Auge					
linkes Auge					

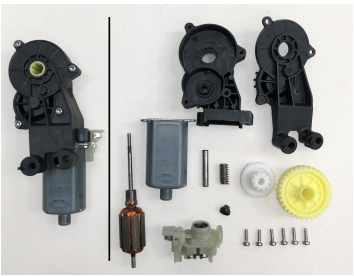


Abb. 6: Verwendeter Elektromotor (Variante A) als ganzes (links) und als demontiertes Produkt (rechts).

Durchgang	Variante	Zeit (Sek.)
1	A	125,72 ± 44,50
2	A	95,03 ± 28,18
3	A	91,53 ± 44,40
4	A	75,70 ± 16,27
5	A	77,25 ± 20,19
6	A	73,25 ± 21,72
7	A' (Defekt)	113,50 ± 35,60
8	B	101,27 ± 33,88
9	C	82,70 ± 30,64
10	A	69,84 ± 20,93

Tab. 2: Darstellung der mittleren Demontagedauer inkl. Standardabweichung von allen Versuchspersonen pro Versuch.

personen teilgenommen. Davon waren 6 weiblich und 9 männlich, mit einem durchschnittlichen Alter von 25,5 (±3,2) Jahren. Die teilnehmenden Personen verfügten über technisches Hintergrundwissen und schätzten die eigene handwerkliche Vorerfahrung in einem abschließenden Fragebogen ein. Die Kategorien bestanden aus *wenig bis keine Vorerfahrung* (n=1), *mittlere Vorerfahrung - der Umgang mit Schraubenzieher, Akkuschrauber, Zange sind bekannt* (n=9), *umfangreiche Vorerfahrung - Hobby-HandwerkerIn* (n=5) und *fachkundige Vorerfahrung - Ausübung eines Handwerksberufs* (n=0). Zusätzlich wurde in diesem Fragebogen offene Fragen zur durchgeführten Demontage gestellt. Der Fokus lag hierbei auf der Beschreibung von eingesetzten Strategien, von komplexen und schwierig zu lösenden Bauteilen sowie von bereits vorhandenem Wissen. Im Zuge des Anwendungsfalls des Remanufacturing wurden den Versuchspersonen verschiedene Varianten eines Elektromotors gemäß Abbildung 6 zur Demon-

tage vorgelegt. Insgesamt gab es zehn Demontagedurchgänge. Vor der ersten Demontage konnten sich die Versuchspersonen an die Plattform und an die Verfolgung durch den Roboterarm gewöhnen. Zudem wurden die eingebauten Sicherheitsmaßnahmen erklärt, um jederzeit die Bewegung des Roboterarms selbstständig zu stoppen. In der gesamte Studie musste kein Durchgang aufgrund von Komplikationen oder Unwohlsein abgebrochen werden. Die Verfolgung durch den Roboterarm wurde als nicht störend beschrieben und es konnte kein Einfluss in der Demontagetätigkeit beobachtet werden. Die generelle Aufgabe bestand aus der kompletten Demontage des Elektromotors. Währenddessen gab es keine zeitliche Begrenzung und zunächst keine Hinweise auf die durchzuführenden Demontageprozesse. In den ersten sechs Durchgängen mussten die Versuchspersonen dieselbe defektfreie Variante A des Elektromotors demontieren. Somit wurde der Demontageprozess trainiert und spezifisches Handlungswissen aufgebaut. Die wesentlichen Prozesse bestanden aus dem Lösen von Schraubverbindungen (Getriebedeckel/-gehäuse, Antriebsschnecke), dem Trennen von Steckverbindungen (Getriebedeckel/-gehäuse, Bürstenhalter) und dem Entnehmen von Bauteilkomponenten (Zahnräder, Rotor). Als besonders komplizierten und nicht direkt erkennbaren Demontageschritt wurde das Abziehen des Bürstenhalters vom Poltopf in dem Fragebogen beschrieben. Häufig wurde dieser Prozess in den ersten zwei Durchgängen nicht erkannt. Im Falle von nicht demontierten Bauteilkomponenten gab es generell nach dem zweiten Durchgang einen Hinweis durch die Studienleitung auf noch weitere Demontageschritte. Dieser einzige Hinweis wiederholte sich zu Beginn der folgenden Durchgängen, bis der Elektromotor komplett demontiert wurde. Spätestens vor dem sechsten Durchgang wurden alle erforderlichen Demontageprozesse aufgelöst. Nach den ersten sechs Durchgängen wurden in randomisierter Reihenfolge die Varianten B und C sowie die zu A gespiegelte Variante A' demontiert. Die Varianten B und C unterschieden sich primär in den Dimensionen des Getriebes. Die gespiegelte Variante A' wies zudem immer einen standardisierten und reproduzierbaren Defekt (verklemmte Bauteilkomponente) auf. Dieses Problem war zunächst nicht direkt von außen sichtbar und trat erst innerhalb des Demontageprozesses auf. Somit wurde eine reale und standardisierte Demontage zusammen mit den komplexen Anforderungen des Remanufacturing ermöglicht. Im letzten Durchgang musste die defektfreie Variante A demontiert werden. Danach erfolgte die Bearbeitung des genannten Fragebogens.

Während der Demontage wurde der Mensch gemäß Abschnitt 3 erfasst. Die gewonnenen Daten bestehen unter anderem aus einem RGB-D-Bildstrom der statischen und der beweglichen RGB-D-Kamera. Die Farbbilder haben eine Auflösung von 1920 px × 1080 px, die Tiefenbilder sind mit 1024 px × 1024 px aufgelöst. Die Bildwiederholrate der Farb- und Tiefenbilder beträgt für beide Kameras

30 Hz. Höhere Auflösungen der Farbbilder sind zulasten der Bildwiederholrate möglich, jedoch erwiesen sich die genannten Auflösungen als ausreichend. Die Messung der Augen- und Blickbewegung erfolgte mit dem kopfgetragenen Blickregistrierungssystem SMI Eye Tracking Glasses 2w. Das binokulare System erfasst die Pupillen der Versuchsperson mit zwei Infrarotkameras und einer Bildwiederholrate von 60 Hz. Zudem wird mit einer Szenenkamera eine Videoaufnahme aus der Ich-Perspektive der Versuchsperson aufgezeichnet.

Anhand der Bilddaten erfolgt die Berechnung der einzelnen Demontagezeiten. Zu Beginn der Demontage wurde jeder Elektromotor in einem Kleinladungsträger der Versuchsperson angereicht. Die erste initiale Handbewegung war das Herausnehmen des Motors und der erste Kontakt zwischen der Hand und dem Elektromotor definiert somit den Startzeitpunkt. Während der Demontage waren die Versuchspersonen angewiesen, die jeweils demontierten Bauteilkomponenten zurück in den Kleinladungsträger abzulegen. Der Endzeitpunkt ist durch den letzten Kontakt zwischen der Hand und der letzten abgelegten Bauteilkomponente gekennzeichnet. In Tabelle 2 sind die Demontagezeiten für jeden Durchgang und für die jeweilige Variante dargestellt. Daran lässt sich eine schnelle Lernkurve in der Demontage der Variante A des Elektromotors ablesen. Bereits im zweiten und drittem Durchgang hat sich die Demontagezeit um etwa 25% im Vergleich zum ersten Durchgang verbessert. Der sechste Durchgang (~40%) und der letzte Durchgang (~45%) waren annähernd zur Hälfte schneller. Daher kann von einem schnellen Training im Umgang mit der Variante A des Elektromotors und der Adaption von demontagerelevantem Wissen ausgegangen werden. Die Variante A' (~10%) hat aufgrund des Defekts ähnlich lange gedauert, wie der erste Demontagedurchgang von der Variante A. Die Variante B (~20%) und die Variante C (~35%) waren wieder deutlich schneller als der erste Demontagedurchgang von Variante A des Elektromotors.

Wie in Unterabschnitt 3.2 erläutert, werden auch einige Daten abgeleitet, wie bspw. die 2D-Pose des Menschen durch OpenPose. In Verbindung mit den Tiefenbildern kann dann die 3D-Pose des Menschen im Arbeitsraum rekonstruiert werden. Weiterhin wird auf den RGB-Bildern der beweglichen RGB-D-Kamera eine Objekterkennung ausgeführt, um Werkzeuge oder Einzelteile erkennen zu können. Die eben beschriebenen Modalitäten (Rohdaten und einige abgeleitete Daten) sind in Abbildung 2 und in Abbildung 4 dargestellt. In Zukunft sollen mittels Tiefeninformationen der beweglichen RGB-D-Kamera und der Vorwärtskinematik des Roboterarms die Objekte im Arbeitsraum lokalisiert werden. Die 3D-Pose des Menschen, dessen Hände, sowie erkannte Objekte und Werkzeuge dienen zukünftig als Eingabe zur zweihändigen Aktionserkennung (Abschnitt 5).

Um eine systematische Vergleichbarkeit maschineller Lernverfahren unterschiedlicher Anwendungsbereiche auf zukünftigen Evaluationen des Datensatzes

sicherzustellen, sowie für die in Abschnitt 6 vorgestellte Clustering-Analyse, wird der Datensatz in einen Trainings-/Validierungsdatensatz und in einen Testdatensatz aufgeteilt. Der Testdatensatz besteht aus spezifisch ausgewählten Versuchspersonen. Darin enthalten sind sowohl zwei weibliche als auch zwei männliche Versuchspersonen (Versuchspersonnummer 12–15) mit jeweils einer mittleren handwerklichen Vorerfahrung als auch mit einer umfangreichen handwerklichen Vorerfahrung. Der gesamte Testdatensatz umfasst 40 Aufnahmen mit einer Gesamtdauer von etwa 75 Minuten und einer durchschnittlichen Aufnahmedauer von $112{,}40 \pm 39{,}96$ Sekunden.

5 Zweihändige Aktionserkennung

In unserer vorherigen Arbeit [9] wurde ein Ansatz zur zweihändigen Aktionserkennung basierend auf neuronalen Graphnetzwerken vorgestellt. Der dort beschriebene Klassifikator nimmt einen Szenengraphen als Eingabe, und gibt als Ausgabe die erkannten Aktionen aus, die die linke und rechte Hand ausführen, wie beispielsweise „hinlangen", „schrauben", „hämmern", aber auch Aktionen aus dem Haushaltskontext. Ein Szenengraph nach [9] ist ein rein symbolisches Modell der aktuellen Szene, wobei Knoten Einzelteile, Hände, oder Werkzeuge repräsentieren, und Kanten die jeweiligen symbolischen räumlichen Relationen zwischen diesen, wie bspw. „in Kontakt", „darüber", „rechts von", usw. Um die Objekte zu erkennen, wurde dabei YOLOv3 [37] verwendet, während zur Erkennung der menschlichen Pose auch OpenPose zum Einsatz kam. Im Szenengraphen als Eingabe für das Graphnetzwerk werden die genannten Informationen durch Knoten- und Kantengewichte repräsentiert. Die Knotengewichte werden dargestellt als $\mathrm{ohe}(\mathrm{id}_{\mathrm{class}})$, wobei ohe die One-Hot-Kodierung der Klassennummer $\mathrm{id}_{\mathrm{class}}$ ist. Die Kantengewichte wurden als ein Vektor $w \in \{0,1\}^{|R|}$ repräsentiert, wobei R die Menge aller möglichen räumlichen Relationen ist. Der Vektor ist genau dann an der i-ten Stelle 1, wenn die i-te räumliche Relation zwischen den Objekten wahr ist, und 0 andernfalls. Durch die Darstellung der Szene als Graph ist unser Ansatz nicht empfindlich im Hinblick auf das Hinzufügen oder Entfernen von Objekten oder veränderten Objektreihenfolgen, die bei der Objekterkennung auftreten können.

In dieser Arbeit wurde der vorgestellte Ansatz aus [9] erweitert, um auch subsymbolische Informationen im Szenengraphen zu repräsentieren. So wurden Objekte hinzugefügt, die die Pose des Menschen beschreiben, sowie eine neue Relation, welche kinematische Zwangsbedingungen zwischen Posenobjekten repräsentiert. Diese Posenobjekte umfassen, von der bereits vorher vorhandenen

linken und rechten Hand abgesehen, auch den linken und rechten Ellenbogen, die linke und rechte Schulter, sowie den Kopf. Die subsymbolischen Informationen, nämlich die Pose des Menschen, wird im Szenengraph dadurch repräsentiert, dass den Knotengewichten, die Posenobjekte repräsentieren, zusätzlich zur One-Hot-Kodierung der Klassennummer ohe($\mathrm{id}_{\mathrm{class}}$) die Position des Posenobjekts (x, y, z) in Form von 3 weiteren Dimensionen hinzugefügt wird. Im Falle von normalen Objekten oder Werkzeugen sind diese Gewichte $(0, 0, 0)$. In Abbildung 7 ist das dargestellt durch Knoten, die mit blau umrandet sind. Dieses Vorgehen ist teilweise vergleichbar mit neueren Arbeiten zur skelettbasierten menschlichen Aktionserkennung, die auch auf Graphnetzwerke setzen [41, 54, 30]. Um das Lernproblem einfach zu halten, wurde eine Objektklasse eingeführt, die Ellenbogen und Schultern zusammenfasst. Dies ist in Abbildung 7 dargestellt durch Knoten, die nicht farbig ausgefüllt sind.Eine Unterscheidung von Ellenbogen und Schultern kann allerdings weiterhin durch die kinematischen Randbedingungen, modelliert durch die neu dafür eingeführte Kantenart, erfolgen. Hierzu wurden die Vektoren w, welche Kantengewichte enthalten, um eine zusätzliche Dimension erweitert: $w \in \{0, 1\}^{|R+1|}$. Räumliche Relationen werden weiterhin wie vorher beschrieben repräsentiert, mit der Ergänzung, dass die letzte Dimension 0 sein muss. Ist ein Kantengewichtsvektor der Form $w = (0, \ldots, 0, 1)$, so wird dadurch eine kinematische Randbedingung zwischen Posenobjekten ausgedrückt. Das heißt, dass eine Kante, welche kinematische Randbedingungen ausdrückt, das Vorhandensein einer Kante, welche räumliche Relationen ausdrückt, ausschließt, und umgekehrt. Eine Visualisierung des erweiterten Szenengraphen, einschließlich der ursprünglichen Repräsentation aus [9] ist in Abbildung 7 dargestellt.

Evaluiert wurde der vorgestellte Klassifikator wie in [9] beschrieben auf dem dort vorgestellten Datensatz. Hierbei wurde eine Leave-One-Out-Kreuzvalidierung durch Anwendung des Klassifikators auf Daten zuvor ungesehener Probanden und Berechnung des Aktionsklassifikations-Macro-f_1-Scores durchgeführt. Durchschnittlich über 6 Probanden erzielte der vorgestellte Klassifikator einen Aktionsklassifikations-Macro-f_1-Score von 0.70 (vgl. zu [9]: 0.63) wenn die Aktion mit der höchsten ausgegebenen Wahrscheinlichkeit der tatsächlich ausgeführten Aktion entspricht. Durch Auflockerung dieser Bedingung, bei der die Ausgabe des Klassifikators als korrekt bewertet wird, wenn die tatsächlich ausgeführte Aktion eine der 3 wahrscheinlichsten Aktionen des Klassifikators ist, wird ein Aktionsklassifikations-Macro-f_1-Score von 0.92 erreicht (vgl. zu [9]: 0.86).

In Zukunft wird die zweihändige Aktionserkennung in den vorgestellten Versuchsaufbau (Abschnitt 3) und unter Verwendung der erhobenen Daten aus Abschnitt 4 integriert werden. Dabei soll auf die Objekterkennung von Detectron2 gesetzt werden, die eine präzisere Modellierung der Objekte durch beliebig aus-

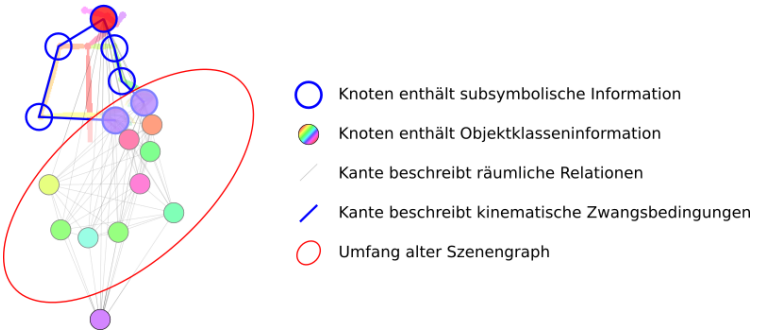


Abb. 7: Grafische Darstellung eines erweiterten Szenengraphen als Eingabe für einen Graphnetzwerk-Klassifikator. Knoten repräsentieren Objekte, Werkzeuge, Hände, oder andere Schlüsselpunkte der menschlichen Pose. Beinhaltet ein Knoten subsymbolische Information (Position im Kamerakoordinatensystem), so ist er dick blau umrandet. Ist die Objektklasse zu einem Knoten im Szenengraphen hinterlegt, so ist der Knoten farblich ausgefüllt. Kanten repräsentieren räumliche Relationen (dünne, graue Kanten) oder kinematische Zwangsbedingungen (dicke, blaue Kanten). Der alte Szenengraphen ohne Tisch (unten) sowie weiterer Schlüsselpunkte der menschlichen Pose ist rot umkreist dargestellt.

gerichtete Quader (Object-Oriented Bounding Boxes) statt achsenausgerichtete Quader (Axis-Aligned Bounding Boxes) ermöglicht, und somit die Genauigkeit der räumlichen Relationsdaten erhöht. Darüber hinaus wird die Hand-Keypoint-Detection des OpenPose-Frameworks genutzt werden, um die Pose der Hände zu erkennen, und somit Kontaktrelationen zwischen den Händen des Menschen und den Objekten zu ermitteln.

6 Clustering-Analyse zur unüberwachten Erkennung von Handlungsabläufen

In diesem Abschnitt führen wir eine qualitative Analyse der erfassten Videoaufnahmen mittels eines unüberwachten Clustering-Verfahrens auf den Daten des in Abschnitt 4 beschriebenen Testdatensatzes durch. Hierbei beschränken wir die Analyse auf die Variante A, um ein ungewolltes Clustering der Daten entlang unterschiedlicher Varianten zu vermeiden. Ziel ist es, einen ersten Überblick über die Eignung der Daten zur Erkennung von Handlungsabläufen zu gewinnen. Dafür unterteilen wir die Aufnahmen in Video-Snippets zu je 32 Frames (ca. 1 Sekunde) und transformieren diese anschließend mit einem Video-Encoder in

Vektorrepräsentationen, welche unter der Verwendung des k-Nearest-Neighbour Clustering partitioniert werden.

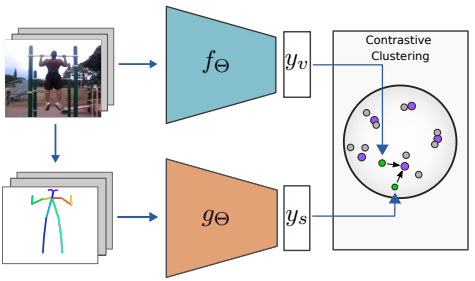


Abb. 8: Der Video-Encoder f_Θ und der Posen-Encoder g_Θ werden darauf trainiert, ähnliche Repräsentationen y_v und y_s zu erzeugen. Nach dem Vortraining können die Encoder dafür genutzt werden, ohne weitere Anpassung auf den Zieldaten nutzbare Repräsentationen zu erzeugen.

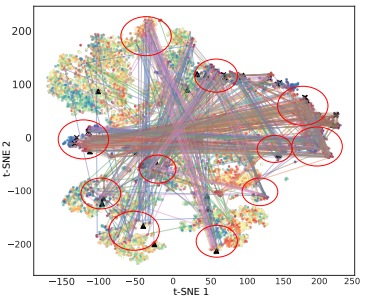


Abb. 9: Darstellung der Aktionsraumpfade innerhalb der ersten 12,5 Sekunden über alle Videos des Testdatensatzes hinweg.

Video-Encoder. Moderne faltende neuronale Netze (engl. Convolutional Neural Networks, CNNs) zur Handlungserkennung in Videoaufnahmen bieten zwar hohe Klassifizierungsgenauigkeiten, jedoch handelt es sich um sehr umfangreiche Architekturen. Als Video-Encoder benutzen wir daher Separable 3D ConvNet (S3D), ein ressourceneffizientes dreidimensionales CNN, welches durch die Trennung von temporalen und räumlichen Filtern in den ersten Layern des Netzes mit einer vergleichsweise geringen Anzahl von Parametern auskommt. Die Architektur baut auf mehreren charakteristischen Inception-Modulen auf – kleinen Sub-Netzwerken, die $1\times3\times3-3\times1\times1$ und $1\times1\times1$ Faltungsoperationen parallel ausführen und die Ausgabe verknüpfen, was die Komplexität des Netzes weiter reduziert. Das komplette Netzwerk besteht aus 27 Schichten.

Der beschriebene S3D Video-Encoder wird mit einem selbstüberwachten Lernverfahren gemäß [40] auf dem großen öffentlich verfügbaren Videodatensatz Kinetics-400 vortrainiert. Aufgrund des verwendeten Ansatzes werden keine Aktions-Annotationen benötigt, wie sie durch Kinetics-400 bereitgestellt werden, damit eignet sich diese Methode in Zukunft auch für ein Vortraining auf deutlich größeren, nicht annotierten Datensätzen. Die Verwendung eines vortrainierten Netzwerkes ermöglicht eine Enkodierung der RGB-Videodaten ohne expliziete Anpassung auf den Zieldaten. Diese Eigenschaft ist bei unserer Analyse des Datensatzes wünschenswert, da in der Praxis gegebenenfalls keine größeren Da-

tensätze für eine Anpassung zur Verfügung stehen, sondern ein Handlungsablauf durch einmalige Demonstration erkannt werden soll.

Das Vortraining erfolgte mittels des in Abbildung 8 dargestellten kontrastierenden Lernverfahrens von [40], welches die Datenmodalitäten RGB-Video und Posensequenz miteinander kombiniert. Hierfür werden auf einem separaten Trainingsdatensatz ohne menschliche Annotationsarbeit Posen durch Anwendung des OpenPose-Frameworks [6] automatisch extrahiert, worauf sie gepaart zu den Videosequenzen vorliegen. Während des Trainings werden ein Video- und ein Posensequenzencoder darauf folgend parallel konditioniert, zueinander ähnliche Repräsentationsvektoren für Teilsequenzen zu erzeugen, welche sich dazu eignen, die Repräsentation der Teilsequenz der jeweils anderen Modalität in einer Menge von Repräsentationsvektoren eindeutig zu identifizieren. Die erzeugten Repräsentationen werden innerhalb eines Batches in jedem Durchlauf einer Menge von Cluster-Prototypen C zugeordnet. Für die Inferenz auf dem Zieldatensatz werden die lediglich für das Vortraining benötigten Cluster-Prototypen verworfen und nur die eigentlichen 256-dimensionalen Repräsentationsvektoren des Video-Encoders verwendet. Das Lernsignal für die Optimierung des S3D-Netzwerkes wird durch die Zuordnung der Repräsentation der Posensequenz erzeugt, die Optimierung erfolgt mit dem Backpropagation-Algorithmus. Da das Vortraining eine Extraktion von Informationen erzwingt, welche sich zur Zuordnung von Video- und Posensequenzen eignen, lernt der Video-Encoder besonderen Fokus auf Bewegungen des menschlichen Körpers zu legen, ist im Gegensatz zu einem rein posenbasierten Encoder aber in der Lage, Kontextinformationen wie die Umgebung oder im Bild vorhandene Objekte zusätzlich zu berücksichtigen.

Clustering Analyse. Die erfassten Video-Repräsentationen wurden mittels k-Means-Clustering 20 unterschiedlichen Aktionsgruppen zugeordnet und anschließend qualitativ analysiert (Abbildung 10). Diese Cluster-Zuordnung ist mit unterschiedlichen Farben markiert, und mittels *t-Stochastic-Neighbourhood-Embedding (t-SNE)* in einer zweidimensionalen Darstellung abgebildet. Die Farbe unterschiedlicher Punkte repräsentiert die jeweilige Cluster ID. Die Gesamtzahl der Cluster wurde manuell mittels qualitativer Analyse auf dem Validierungsdatensatz auf 20 gesetzt. Qualitative Analysen ergeben, dass ähnliche Handlungen ähnlichen Clustern zugeordnet werden (z.B. die Nutzung eines Akkuschraubers im mit der Farbe Hell-Lila markierten Cluster oben links in der Abbildung 10). Allerdings beobachten wir auch gewisse Biases, bspw. in Bezug auf die Erscheinung der Versuchsperson. Zum Beispiel, beobachten wir, dass Beispiele mit Nutzung des Akkuschraubers öfters im zweidimensionalen Raum (Abbildung 10) voneinander getrennt sind. Durch das Vortraining mit Fokus auf Körperbewegungen wird dieser erscheinungsbasierte Bias gezielt verringert. In Abbildung 11 sind die zeitlichen Verläufe von drei aufgenommenen Versuchen dargestellt. Die Zeit

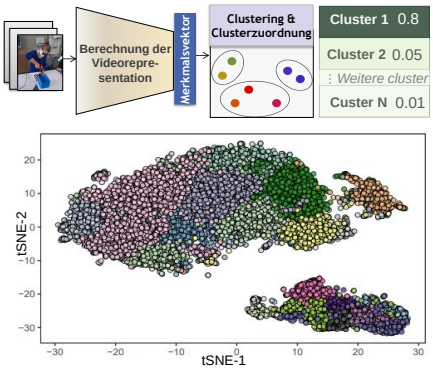


Abb. 10: Clustering-Analyse der Videorepräsentationen im aufgenommenen Datensatz. Video-Snippets der ungefähren Dauer von einer Sekunde werden durch ein mit Self-supervised Learning vortrainiertes Modell in einen Vektor transformiert und anschließend geclustert (Farbe = Cluster ID).

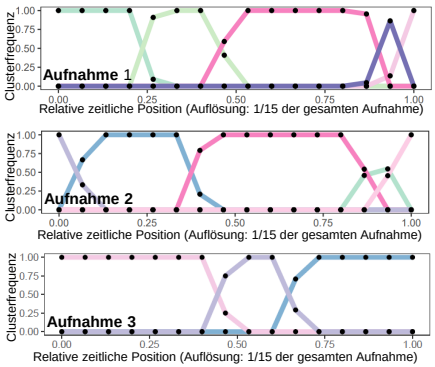


Abb. 11: Clusterzuordnungen in drei verschiedenen Versuchsverläufen. Die Farbe repräsentiert die Cluster-ID, die Zeit ist auf der X-Achse abgebildet, die Y-Achse zeigt den Prozentsatz der Daten, die während eines Zeitschritts ($\frac{1}{15}$ einer Aufnahme) dem jeweiligen Cluster zugeordnet wurden auf.

ist auf der X-Achse abgebildet, die Y-Achse zeigt den Prozentsatz der Daten auf, die während eines Zeitschritts ($\frac{1}{15}$ einer Aufnahme) dem jeweiligen Cluster zugeordnet wurden. Die resultierenden Cluster (Linien mit unterschiedlichen Farben) spiegeln die unterschiedlichen Phasen des Versuchsablaufs wieder.

In Abbildung 9 werden die Video-Snippets weiter auf ihren zeitlichen Verlauf innerhalb des enkodierten Repräsentationsraumes untersucht, wobei jedes Video des Datensatzes einem eigenen gezeichneten Pfad entspricht. Insbesondere in den gezeigten ersten 12,5 Sekunden einer gegebenen Aufgabe sind klare Ablaufmuster im Repräsentationsraum über alle Videos hinweg zu erkennen. Videos gruppieren sich nicht anhand oberflächlicher Merkmale wie bspw. der Identität oder Kleidung der handelnden Person, was zu einer Bewegung der Ablaufpfade innerhalb eines abgegrenzten Bereiches führen würde. Stattdessen decken alle Videos im zeitlichen Verlauf ähnliche Bereiche im Repräsentationsraum ab. Mutmaßlich ist dies auf die identische Aufgabenstellung und die damit einhergehende ähnliche Handlungsabfolge in jedem Video zurückzuführen. Dies ist die erste Validierung, dass die extrahierten Repräsentationen über ein einzelnes Video hinausgehende und interpretierbare Informationen erfassen.

7 Externalisierung von implizitem menschlichem Handlungswissen

Das implizite Handlungswissen des Menschen besteht aus Wissen zum erfahrungsabhängigen und problembezogenen Handeln. Es wird individuell aufgebaut, ist situations- oder kontextabhängig und integriert Informationen aus der komplexen sensorischen Wahrnehmung des Menschen [19]. Dabei bewirkt die unzertrennbare Wechselwirkung zwischen Kognition, Sensorik und Motorik einen effektiven Lernprozess [26]. Neben der Speicherung von bereits ausgeführten Handlungen und Problemlösestrategien werden auch Informationen über Objekteigenschaften und Bewegungsausführungen erfasst. Der Mensch besitzt somit die Fähigkeit sein umfassendes implizites Handlungswissen zu reproduzieren und gezielt zur Bewältigung von neuen Aufgaben oder adaptiv zur Problemlösung einzusetzen [15]. Dieses – teils unbewusst vorliegende – Wissen ist nicht nur schwer zugänglich, sondern kann auch nur teilweise artikuliert werden.

Die Analyse der menschlichen Augen- und Blickbewegung als ein valides und etabliertes Verfahren zur Erfassung kognitiver Prozesse [10, 21] stellt jedoch eine geeignete Methode dar, um dieses implizite Wissen des Menschen zu externalisieren. Es wird bereits zur Beschreibung von internalen Prozessen der Informationsverarbeitung [49], von Problemlösestrategien [27, 14] und der selektiven Aufmerksamkeit [34, 45] verwendet. Bisherige Forschungsarbeiten konnten zeigen, dass mittels der Analyse der visuellen Aufmerksamkeit verbunden mit der selektiven Auswahl von relevanten Umgebungsinformationen, die impliziten Prozesse des Arbeitsgedächtnisses untersucht sowie externalisiert werden können [44]. Weitere Untersuchungen zur Erfassung von Handlungs- und Problemlösestrategien mittels der Blickanalyse erzielten gleichwertige Erkenntnisse. Die Auswertung des Blickpfads ergab die Abfolge der verarbeiteten Reize sowie die Übergänge zwischen informationsrelevanten Interessengebieten [11, 21]. Dieses Blickverhalten kann zur Detektion und Analyse von implizit geplanten Handlungsstrategien verwendet werden. Diese Ergebnisse werden durch empirische Studien zur Untersuchung von Problemlösestrategien in unterschiedlichen Domänen unterstützt. Es konnten Unterschiede im Blickverhalten von Experten und Novizen aufgezeigt werden [3, 20, 50]. Die Experten fixierten schneller, länger und häufiger aufgabenrelevante Bereiche sowie spezifische Informationen zur Problemlösung. Ein weiterer Forschungsbereich ist die Messung und Bestimmung der mentalen Beanspruchung des Menschen. In kontrollierten Versuchsbedingungen wurden unterschiedlich starke mentale Beanspruchungen erzeugt und währenddessen die Augen- und Blickbewegungsparameter der Versuchspersonen erfasst. Es wurde festgestellt, dass aufgrund der kognitiven Prozesse sich spezifische Parameter

signifikant verändern. Unter anderem konnte eine Veränderung des Pupillendurchmessers [29, 43], des Nächsten-Nachbar-Index der gemessenen Fixationen [8] als auch eine erhöhte Blinzelrate [39] beobachtet werden. Diese Reaktionen sind geeignete Indikatoren für die Erfassung der mentalen Aktivität und können dazu beitragen, den Zeitpunkt eines Problems innerhalb einer Handlung festzustellen. Zudem kann beschrieben werden, wie das Problem erkannt und ab welchem Zeitpunkt eine Lösungsstrategie geplant sowie durchgeführt wurde. Des Weiteren untersuchten Studien, ob mittels der erfassten Blickbewegung von Experten das implizite Wissen auf eine andere Person transferiert werden kann [48, 13]. Es zeigte sich, dass eine zuvor betrachtete visualisierte Blickbewegung des Experten einen Effekt auf den Lernprozess von Novizen hatte und dadurch die Leistung im Problemlösen fördern konnte. Dieser gezeigte Wissenstransfer zwischen Mensch und Mensch legt die Existenz einer geeigneten Methode zur Übertragung von implizitem Handlungswissen zwischen Mensch und Maschine nahe. Insgesamt verdeutlichen die Ergebnisse, dass die Analyse der menschlichen Augen- und Blickbewegung ein effektives Verfahren ist, um die kognitiven Prozesse zum Aufbau von implizitem Handlungswissen zu erfassen und zu externalisieren.

Jedoch wurde eine Vielzahl der gefundenen Ergebnisse unter kontrollierten Laborbedingungen erhoben [10]. Es kamen zumeist computerbasierte Aufgaben und standardisierte Testverfahren zum Einsatz. Daher gibt es nur wenige Erkenntnisse zur Übertragbarkeit der gewonnenen Ergebnisse in praxisnahen Anwendungen und insbesondere in eine komplexe Produktionsumgebung mit ungewissen bzw. wandlungsfähigen Situationen. Für eine erste Beurteilung wurde in einer explorativen Machbarkeitsstudie [55] die Augen- und Blickbewegung in einem industriellen Arbeitsumfeld untersucht. Die Untersuchungsaufgabe bestand aus der Demontage eines industriellen Produkts, ein elektrischer Motor aus einem Auto. Dieser Elektromotor wurde mehrfach demontiert und zur Analyse von Handlungs- und Problemlösestrategien erfolgte die Demontage sowohl mit unbeschadeten als auch mit einem defekten Motor. Vergleichbar zu den bisherigen Forschungsergebnissen konnte gezeigt werden, dass während der Durchführung aufgabenrelevante Objekte länger und häufiger fixiert werden. Es konnte eine wahrscheinlichere Blickbewegung innerhalb der informativen sowie der benachbarten Objekte beobachtet werden. Die gewonnenen Erkenntnisse zeigen, dass mittels der Augen- und Blickbewegung es möglich ist, handlungsorientiertes Wissen des Menschen ebenso in einer praxisnahen Anwendung zu erfassen.

Neben der Schwierigkeit, dass etablierte Verfahren der Analyse der Augen- und Blickbewegung in eine wandlungsfähige Produktion zu integrieren, besteht eine weitere Herausforderung im Wissenstransfer zwischen Mensch und Maschine. Bisher ist der Transfer des impliziten Handlungswissens des Menschen zur Nutzung in maschinellen Lernverfahren noch wenig untersucht [38]. Zwar

kann der Mensch, aus der visualisierten Blickbewegung eines Experten, Wissen erlernen [48, 13] aber es fehlen umfangreiche und fundierte Ergebnisse, um auch autonome Systeme, wie Roboter, aus der menschlichen Augen- und Blickbewegung lernen zu lassen. Dennoch gibt es Hinweise in Forschungsarbeiten darauf, dass mittels der Analyse der menschlichen Augen- und Blickbewegung eine lernende Aktionserkennung verbessert und die Programmierung von Robotern durch Demonstrationen unterstützt werden kann [12, 4]. Die bisherigen Ergebnisse basieren auf alltäglichen Aufgaben unter Laborbedingungen und es fehlen weiterhin Erkenntnisse zur Integration in eine komplexe Produktionsumgebung. Zudem muss ein bekanntes Problem in der Auswertung und Interaktion mit Augen- und Blickbewegung, das Midas-Touch-Problem [22], berücksichtigt werden. Im Kontext der Analyse von Augen- und Blickbewegung beschreibt dieses Problem, dass nicht jeder Blick die gleiche Intention oder Bedeutung besitzt. Während einer Handlung können sowohl Blicke ohne eine tiefe kognitive Verarbeitung als auch Blicke zur Betrachtung der Umgebung auftreten. Damit nicht jeder gemessene Blickpunkt eine unwillentliche Handlung oder Aktion im Wissenstransfer zwischen Mensch und Maschine beschreibt, muss zuvor eine zielführende Analyse stattfinden.

Dazu wurde in der beschriebenen Forschungsstudie (siehe Abschnitt 4) die menschliche Augen- und Blickbewegung während der Demontage von Elektromotoren mit einem kopfgetragenen Blickregistrierungssystem erfasst. Ziel ist es, mittels der Blickanalyse Handlungsstrategien zur Demontage und zur Lösung des auftretenden Problems zu beschreiben, sowie den internen kognitiven Zustand des Menschen zu erfassen. Derzeit werden die erhobenen Daten analysiert und mögliche Verfahren zur Verbesserung der beschriebenen zweihändigen Aktionserkennung untersucht. Erste Erkenntnisse zeigen, dass durch den gemessenen Blickpfad auf eine objektbezogene Demontagestrategie geschlossen und relevante Objekte innerhalb der Demontage erkannt werden können. Zudem verdeutlicht die erste Auswertung von Augenbewegungen, dass eine Abweichung oder ein unvorhersehbares Ereignis (z. B. ein Problem) im Demontageprozess identifizierbar ist. Somit können Handlungs- und Problemlösestrategien sowie besondere Schlüsselpunkte in der Demontage festgestellt werden. Die lernende Aktionserkennung wird durch diese Informationen und durch die weitere Modalität der menschlichen Augen- und Blickbewegung ergänzt und soll zu besseren Ergebnissen führen. Zudem basiert die Auswertung auf den in der Literatur beschriebenen Augen- und Blickbewegungsparametern und diese dienen als Anhaltspunkte zur Entwicklung der stationären Blickrichtungsschätzung. Zukünftig sollen die aussagekräftigsten Parameter mit dem in Unterabschnitt 3.3 erläuterten stationären Blickregistrierungssystem messbar sein. Insgesamt soll durch die Forschungsstudie dadurch

ein tiefes Verständnis für das implizite menschliche Handlungswissen und der Externalisierung mittels der Augen- und Blickbewegung aufgebaut werden.

8 Zusammenfassung und Ausblick

In dieser Arbeit wurde eine multisensorielle, robotergestützte Plattform zur Erfassung menschlicher Demontagedemonstrationen vorgestellt. Die Plattform ist durch eine stationäre RGB-D-Kamera dazu in der Lage, eine Erfassung der dreidimensionalen Pose des Menschen im Arbeitsraum durchzuführen, sowie eine bewegliche RGB-D-Kamera durch einen Roboterarm gezielt auf die Hände des Menschen zur Erfassung von Manipulationsaufgaben auszurichten. Außerdem wurden Herausforderungen bei der Integration eines stationären Blickregistrierungssystems dargelegt und entsprechende Lösungen vorgeschlagen. Mit der vorgestellten Forschungsstudie wurden multimodale Daten mehrerer Versuchspersonen aufgezeichnet und ausgewertet.

In Zukunft soll der Klassifikator der zweihändigen Aktionserkennung aus [9] um einen Attention-Mechanismus erweitert werden, der auf der tatsächlichen Aufmerksamkeit des Menschen basiert und durch die geschätzte Blickrichtung modelliert wird. Dazu kann basierend auf der Blickrichtung und deren Schätzunsicherheit die Wahrscheinlichkeit der Betrachtung für alle detektierten Objekte modelliert werden. Eine umfängliche, systematische Evaluierung zur Fusion aller Daten sowie, welche Unsicherheiten mit der geschätzten Blickrichtung einhergehen, steht allerdings noch aus und soll in der Zukunft erfolgen. Durch eine Segmentierung der Demonstrationen durch erkannte Aktionen sollen Modelle erzeugt werden, die sowohl symbolisch, als auch subsymbolisch Aufgaben beschreiben. Diese Modelle sollen benutzt werden, um das Beobachtete zur robotischen Demontage zu reproduzieren. Eine Verfeinerung des Gelernten aus der Beobachtung kann durch Methoden des robotergestützen kinästhetischen Lehrens erfolgen, wie bspw. vorgeschlagen in unserer Arbeit [25].

Da reale Anwendungen in einer Fabrikumgebung in der Regel *dynamisch* sind und neue Arten von Handlungsabfolgen oder Sensorplatzierungen jederzeit aufkommen können, ist neben der Identifizierung der bekannten Situationen auch die Erkennung unbekannter Abläufe ein wichtiger Bestandteil unserer zukünftigen Arbeit. Dadurch sollen bisher unbekannte Situationen, Handlungsabläufe und Objekte gesondert behandelt und idealerweise mit wenig zusätzlichen annotierten Trainingsdaten inkrementell eingelernt werden. Ein langfristiges Ziel ist daher, Verfahren zu entwickeln, die 1.) auf den in Abschnitt 5 und 6 vorgestellten Methoden und Erkenntnissen aufbauen um die aktuelle Aktion des

Menschen automatisch zu erfassen und als interpretierbares atomares Handlungselement symbolisch darzustellen, 2.) bei neuen Daten zwischen bekannten und unbekannten Situationen unterscheiden können, und 3.) auf veränderten oder neuen Handlungsabfolgen gut generalisieren können, ohne dabei eine hohe Menge annotierter Trainingsdaten zu benötigen.

Die Extraktion von semantischen Informationen ist ein Grundbaustein für die Handlungserkennung und -bewertung und kann durch entsprechendes Vortraining auf großen unspezifischen Datensätzen erlernt werden. Durch einen Fokus auf semantische Repräsentationen von Teilaspekten wie bspw. einem Teil des Körpers, Objektinteraktionen oder dem Kontext der Aktion planen wir eine verbesserte Identifizierbarkeit von bekannten und unbekannten Handlungen.

Funding: Das Projekt wird durch die Carl-Zeiss-Stiftung gefördert.

Literatur

[1] E. E. Aksoy, M. Tamosiunaite, and F. Wörgötter. Model-Free Incremental Learning of the Semantics of Manipulation Actions. *Robotics and Autonomous Systems (RAS)*, 71:118–133, Sept. 2015.

[2] N. I. Badler. *Temporal Scene Analysis: Conceptual Descriptions of Object Movements*. PhD thesis, University of Toronto, Toronto, ON, Canada, Feb. 1975.

[3] R. Bednarik. Expertise-dependent visual attention strategies develop over time during debugging with multiple code representations. *International Journal of Human-Computer Studies*, 70:143–155, 2012.

[4] A. Belardinelli and F. Pirri. Bottom-up gaze shifts and fixations learning by imitation. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 37(2):256–271, 2007.

[5] A. Billard, S. Calinon, R. Dillmann, and S. Schaal. Robot Programming by Demonstration. In B. Siciliano and O. Khatib, editors, *Handbook of Robotics*, pages 1371–1394. Springer, 2008.

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, Jan. 2021.

[7] W.-C. Chang. Robotic Assembly of Smartphone Back Shells with Eye-in-Hand Visual Servoing. *Robotics and Computer-Integrated Manufacturing*, 50:102–113, Apr. 2018.

[8] F. Di Nocera, S. Mastrangelo, S. P. Colonna, A. Steinhage, M. Baldauf, and A. Kataria. Mental workload assessment using eye-tracking glasses in a simulated maritime scenario. In *Proceedings of the Human Factors and Ergonomics Society Europe*, pages 14–16, 2015.

[9] C. R. G. Dreher, M. Wächter, and T. Asfour. Learning Object-Action Relations from Bimanual Human Demonstration Using Graph Networks. *Robotics and Automation Letters (RA-L)*, 5(1):187–194, Jan. 2020.

[10] A. T. Duchowski. Gaze-based interaction: A 30 year retrospective. *Computers and Graphics*, 2018.

[11] S. Eraslan, Y. Yesilada, and S. Harper. Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. *Journal of Eye Movements Research*, 9(1):1–19, 2016.

[12] A. H. Fathaliyan, X. Wangt, and V. J. Santos. Exploiting three-dimensional gaze tracking for action recognition during bimanual manipulation to enhance human-robot collaboration. *Frontiers in Robotics and AI*, 5(25):1–15, 2018.

[13] A. Gegenfurtner and M. Seppänen. Transfer of expertise: An eye tracking and think aloud study using dynamic medical visualizations. *Computer and Education*, 63:393–403, 2013.

[14] E. R. Grant and M. J. Spivey. Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, 14(5):462–466, 2003.

[15] R. Grossman and E. Salas. The transfer of training: what really matters. *International Journal of Training and Development*, 15(2):103–120, 2011.

[16] E. D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, 2006.

[17] T. Hamabe, H. Goto, and J. Miura. A Programming by Demonstration System for Human-Robot Collaborative Assembly Tasks. In *International Conference on Robotics and Biomimetics (ROBIO)*, pages 1195–1201, Dec. 2015.

[18] D. W. Hansen and Q. Ji. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 03 2010.

[19] B. Herbig and A. Büssing. Implizites Wissen und erfahrungsgeleitetes Arbeitshandeln: Perspektiven für Arbeit und Organisation. *Arbeit*, 12(1):36–53, 2003.

[20] T. L. Hodgson, A. Bajwa, A. M. Owen, and C. Kennard. The strategic control of gaze direction in the tower of london task. *Journal of Cognitive Neuroscience*, 12(5):894–907, 2000.

[21] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and

J. van de Weijer. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. OUP Oxford, 2011.

[22] R. J. K. Jacob. What you look at is what you get. *Computer*, 26(7):65–66, 1993.

[23] A. Jamal, V. P. Namboodiri, D. Deodhare, and K. Venkatesh. Deep domain adaptation in action space. In *BMVC*, volume 2, page 5, 2018.

[24] S. Jenni, G. Meishvili, and P. Favaro. Video representation learning by recognizing temporal transformations. In *European Conference on Computer Vision*, pages 425–442. Springer, 2020.

[25] C. Klas, F. Hundhausen, J. Gao, C. R. G. Dreher, S. Reither, Y. Zhou, and T. Asfour. The KIT Gripper: A Multi-Functional Gripper for Disassembly Tasks. In *International Conference on Robotics and Automation (ICRA)*, pages 715–721, Xi'an, China, May 2021. IEEE.

[26] G. Knoblich, S. Butterfill, and N. Sebanz. Psychological research on joint action: Theory and data. In B. Ross, editor, *The Psychology of Learning and Motivation*, pages 59–101. Burlington: Academic Press, 2011.

[27] G. Knoblich, S. Ohlsson, and G. E. Raney. An eye movement study of insight problem solving. *Memory and Cognition*, 29(7):1000–1009, 2001.

[28] H. S. Koppula and A. Saxena. Anticipating Human Activities Using Object Affordances for Reactive Robotic Response. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(1):14–29, Jan. 2016.

[29] B. Laeng, S. Sirois, and G. Gredebäck. Pupillometry: A window to the preconscious? *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, 7(1):18–27, 2012.

[30] S. Li, J. Yi, Y. A. Farha, and J. Gall. Pose Refinement Graph Convolutional Network for Skeleton-Based Action Recognition. *IEEE Robotics and Automation Letters*, 6(2):1028–1035, Apr. 2021.

[31] V. Lippiello, B. Siciliano, and L. Villani. Eye-in-Hand/Eye-to-Hand Multi-Camera Visual Servoing. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 5354–5359, Dec. 2005.

[32] A. Muis and K. Ohnishi. Eye-to-Hand Approach on Eye-in-Hand Configuration Within Real-Time Visual Servoing. *IEEE/ASME Transactions on Mechatronics*, 10(4):404–410, Aug. 2005.

[33] S. Parsa and M. Saadat. Human-Robot Collaboration Disassembly Planning for End-of-Life Product Disassembly Process. *Robotics and Computer-Integrated Manufacturing*, 71:102170, Oct. 2021.

[34] M. I. Posner and Y. Cohen. Components of visual orienting. In H. Bouma and D. G. Bouwhuis, editors, *Attention and performance X: Control of language processes*, page 531–556. Hillsdale, NJ: Lawrence Erlbaum, 1984.

[35] A. Priyoni, W. Ijomah, and U. Bititci. Disassembly for remanufacturing: A

systematic literature review, new model development and future research need. *Journal of Industrial Engineering and Management*, 9(4):899–932, 2016.

[36] D. Rakita, B. Mutlu, and M. Gleicher. An Autonomous Dynamic Camera Method for Effective Remote Teleoperation. In *International Conference on Human-Robot Interaction (HRI)*, pages 325–333, Mar. 2018.

[37] J. Redmon and A. Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, Apr. 2018.

[38] N. Rußwinkel. Antizipierende interaktiv lernende autonome agenten. In H.-J. Buxbaum, editor, *Mensch-Roboter-Kollaboration*. Wiesbaden: Springer Fachmedien, ein Teil von Springer Nature 2020, 2020.

[39] S. W. Savage, D. D. Potter, and B. W. Tatler. Does preoccupation impair hazard perception? a simultaneous eeg and eye tracking study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 17:52–62, 2013.

[40] D. Schneider, S. Sarfraz, A. Roitberg, and R. Stiefelhagen. Pose-based contrastive learning for domain agnostic activity representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022.

[41] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan. Skeleton-Based Action Recognition with Hierarchical Spatial Reasoning and Temporal Stack Learning Network. *Pattern Recognition*, 107:107511, Nov. 2020.

[42] R. Siegfried, B. Aminian, and J.-M. Odobez. Manigaze: A dataset for evaluating remote gaze estimator in object manipulation situations. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Short Papers, New York, NY, USA, 2020. Association for Computing Machinery.

[43] S. Sirois and J. Brisson. Pupillometry. wiley interdisciplinary reviews. *Cognitive Science*, 5(6):679–692, 2014.

[44] J. Theeuwes, B. Artem, and C. N. L. Olivers. Interactions between working memory, attention and eye movements. *Acta Psychologica*, 132(2009):106–114, 2009.

[45] L. E. Thomas and A. Lleras. Moving eyes and moving thought: On the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin and Review*, 14(4):663–668, 2007.

[46] M. Toering, I. Gatopoulos, M. Stol, and V. T. Hu. Self-supervised video representation learning with cross-stream prototypical contrasting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 108–118, 2022.

[47] N. Vahrenkamp, M. Wächter, M. Kröhnert, K. Welke, and T. Asfour. The Robot Software Framework ArmarX. *it – Information Technology*, 57(2):99–111, Mar. 2015.

[48] T. van Gog, H. Jarodzka, K. Scheiter, P. Gerjets, and F. Paas. Attention guidance during example study via the model's eye movements. *Computers in Human Behaviour*, 25:785–791, 2009.

[49] B. M. Velichkovsky. Heterarchy of cognition: The depths and the highs of a framework for memory research. *Memory*, 10(5-6):405–419, 2002.

[50] J. N. Vickers. Perception, cognition, and decision training. the quiet eye in action. *Human Kinetics*, 2007.

[51] A. Villanueva and R. Cabeza. Models for gaze tracking systems. *EURASIP Journal on Image and Video Processing*, 2007(1):023570, 2007.

[52] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2, 2019.

[53] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019.

[54] F. Ye, S. Pu, Q. Zhong, C. Li, D. Xie, and H. Tang. Dynamic GCN: Context-Enriched Topology Learning for Skeleton-Based Action Recognition. *arXiv:2007.14690 [cs]*, July 2020.

[55] M. Zaremski and B. Deml. Analyse von Augen- und Blickbewegungen zur Beschreibung von Handlungswissen in der manuellen Demontage. In *Tagungsband 66. GfA-Frühjahrskongress Digitaler Wandel, digitale Arbeit, digitaler Mensch?*, page Beitrag C.6.3. Gfa, Dortmund, 2020.

[56] X. Zhang, Y. Sugano, and A. Bulling. Evaluation of Appearance-Based Methods and Implications for Gaze-Based Applications. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 05 2019.

[57] F. Ziaeetabar, T. Kulvicius, M. Tamosiunaite, and F. Wörgötter. Recognition and Prediction of Manipulation Actions Using Enriched Semantic Event Chains. *Robotics and Autonomous Systems (RAS)*, 110:173–188, Dec. 2018.

Christian R. G. Dreher studierte Informatik am Karlsruher Institut für Technologie (KIT), und schloss 2019 sein Studium ab. Derzeit arbeitet er als wissenschaftlicher Mitarbeiter am Lehrstuhl für Hochperformante Humanoide Technologien (H2T), KIT. Sein Forschungsinteresse umfasst die Roboterprogrammierung durch Vormachen. **Adresse**: M. Sc. Christian R. G. Dreher • Karlsruher Institut für Technologie (KIT) • Institut für Anthropomatik und Robotik (IAR) • Hochperformante Humanoide Technologien (H2T) • Adenauerring 2 • 76133 Karlsruhe • Deutschland • c.dreher@kit.edu

Manuel Zaremski schloss sein Studium der Bewegungswissenschaft an der Justus-Liebig-Universität Gießen (JLU) 2017 ab und ist wissenschaftlicher Mitarbeiter am Institut für Arbeitswissenschaft und Betriebsorganisation am KIT. Seine Forschungsinteressen liegen im Bereich der Mensch-Maschine-Interaktion und dort beschäftigt er sich insbesondere mit der Analyse der menschlichen Augen- und Blickbewegung. **Adresse**: M. Sc. Manuel Zaremski • Karlsruher Institut für Technologie (KIT) • Institut für Arbeitswissenschaft und Betriebsorganisation (ifab) • Engler-Bunte-Ring 4 • 76131 Karlsruhe • Deutschland • manuel.zaremski@kit.edu

Fabian Leven schloss sein Studium der Physik am Karlsruher Institut für Technologie (KIT) 2019 ab. Momentan ist er wissenschaftlicher Mitarbeiter am Institut für Industrielle Informatiostechnik am KIT. Seine Forschungsinteressen liegen im Bereich des maschinellen Sehens und dort beschäftigt er sich insbesondere mit der Schätzung der menschlichen Blickrichtung. **Adresse**: M. Sc. Fabian Leven • Karlsruher Institut für Technologie (KIT) • Institut für Industrielle Informationstechnik (IIIT) • Hertzstraße 16 • 76187 Karlsruhe • Deutschland • fabian.leven@kit.edu

David Schneider schloss seinen M.Sc. am Karlsruher Institut für Technologie (KIT) ab und ist wissenschaftlicher Mitarbeiter im Computer Vision for Human-Computer Interaction Lab am KIT. Er arbeitet an der Erkennung menschlicher Aktivitäten als Bestandteil assistiver Technologien. **Adresse**: David Schneider • Karlsruher Institut für Technologie (KIT)

• Institut für Anthropomatik und Robotik (IAR) • Vincenz-Priessnitz-Str. 3 • 76131 Karlsruhe • Deutschland • david.schneider@kit.edu

Alina Roitberg Alina Roitberg schloss ihre Promotion 2021 am KIT ab, wofür sie den Dissertationspreis der IEEE ITSS Gesellschaft und den Promotionspreis des KIT erhielt. Ihre Forschung zielt darauf ab, zuverlässige, interpretierbare und dateneffiziente Algorithmen zur Aktivitätenerkennung zu entwickeln. **Adresse**: Dr.-Ing. Alina Roitberg • Karlsruher Institut für Technologie (KIT) • Institut für Anthropomatik und Robotik (IAR) • Vincenz-Priessnitz-Str. 3 • 76131 Karlsruhe • Deutschland • alina.roitberg@kit.edu

Rainer Stiefelhagen ist Professor für Informationstechnische Systeme für sehbeeinträchtigte Studierende am KIT, wo er das Computer Vision for Human-Computer Interaction Lab am Institut für Anthropomatik und Robotik leitet. Er befasst sich mit visueller Erkennung in Bildern und Videos. **Adresse**: Prof. Dr.-Ing. Rainer Stiefelhagen • Karlsruher Institut für Technologie (KIT) • Institut für Anthropomatik und Robotik (IAR) • Vincenz-Priessnitz-Str. 3 • 76131 Karlsruhe • Deutschland • rainer.stiefelhagen@kit.edu

Michael Heizmann ist Professor für Mechatronische Messsysteme und Institutsleiter am Institut für Industrielle Informationstechnik (IIIT) des Karlsruher Instituts für Technologie (KIT). Seine Forschungsgebiete umfassen automatische Sichtprüfung, Signal- und Bildverarbeitung, Bild- und Informationsfusion, Messtechnik, maschinelles Lernen und künstliche Intelligenz sowie deren Anwendungen. **Adresse**: Prof. Dr.-Ing. Michael Heizmann • Karlsruher Institut für Technologie (KIT) • Institut für Industrielle Informationstechnik (IIIT) • Hertzstraße 16 • 76187 Karlsruhe • Deutschland • michael.heizmann@kit.edu

Barbara Deml ist Professorin für Arbeitswissenschaft und Institutsleiterin am Institut für Arbeitswissenschaft und Betriebsorganisation (ifab) des Karlsruher Instituts für Technologie (KIT). Die Forschungsgebiete umfassen die empirische

Analyse des menschlichen Verhaltens und die damit assoziierten kognitiven Prozesse, die Interaktion zwischen Mensch und Maschine sowie lernender automatisierter Systeme und die menschzentrierte Gestaltung von Arbeitssystemen.
Adresse: Prof. Dr.-Ing. Dipl-Psych. Barbara Deml • Karlsruher Institut für Technologie (KIT) • Institut für Arbeitswissenschaft und Betriebsorganisation (ifab) • Engler-Bunte-Ring 4 • 76131 Karlsruhe • Deutschland • barbara.deml@kit.edu

Tamim Asfour ist Professor für Humanoide Robotiksysteme am Karlsruher Institut für Technologie (KIT), wo er den Lehrstuhl für Hochperformante Humanoide Technologien (H2T) leitet. Sein Forschungsinteresse richtet sich an humanoide Roboter, welche aus Beobachtung und Erfahrung lernen können, und in realen Umgebungen agieren und interagieren können.
Adresse: Prof. Dr.-Ing. Tamim Asfour • Karlsruher Institut für Technologie (KIT) • Institut für Anthropomatik und Robotik (IAR) • Hochperformante Humanoide Technologien (H2T) • Adenauerring 2 • 76133 Karlsruhe • Deutschland • asfour@kit.edu