TECHNISCHE UNIVERSITÄT
CHEMNITZ

Fakultät für Informatik

CSR-20-02

Schlussbericht zum InnoProfile-Transfer Begleitprojekt localizeIT

Robert Manthey · Tom Kretzschmar · Falk Schmidsberger · Hussein Hussein · René Erler · Tobias Schlosser · Frederik Beuth · Marcel Heinz · Thomas Kronfeld · Maximilian Eibl · Marc Ritter · Danny Kowerko

Januar 2020

Chemnitzer Informatik-Berichte

Schlussbericht zum InnoProfile-Transfer Begleitprojekt localizeIT

Ansprechpartner:
Jun.-Prof. Dr. rer. nat. Danny Kowerko
Technische Universität Chemnitz
Stiftungsjuniorprofessur Media Computing
Straße der Nationen 62
09111 Chemnitz

Zuwendungsempfänger:
TU Chemnitz
Fakultät für Informatik
Professur für Medieninformatik
Stiftungsjuniorprofessur Media Computing

Förderkennzeichen:
03IPT608X

Vorhabenbezeichnung:
localizeIT — Lokalisierung visueller Medien

Laufzeit des Vorhabens:
01.08.2014 bis 31.07.2019

Berichtszeitraum:
01.08.2014 bis 31.07.2019

Abgabedatum:
31.01.2020

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren:
Robert Manthey, Tom Kretzschmar, Falk Schmidsberger, Hussein Hussein, René Erler, Tobias Schlosser, Frederik Beuth, Marcel Heinz, Thomas Kronfeld, Danny Kowerko (Projektleiter), Marc Ritter (Projektleiter), Maximilian Eibl (Projektinitiator)

GEFÖRDERT VOM

Inhaltsverzeichnis

1 Kurzdarstellung

1.1 Aufgabenstellung

Das Akronym LocalizeIT ist bewusst doppeldeutig gewählt. Die Übersetzung „Lokalisiere es“ adressiert die im Projektfokus stehende Lokalisierung. Hinter dem „es“ verbergen sich verschiedenste Objekte und Personen. Zu den Objekten zählen neben sämtlichen Inhalten digitaler Fotografien des Alltags auch Objekte industrieller Fertigungsprozesse wie etwa in Laser-verarbeitenden Prozessen im Bereich Maschinenbau. Die Buchstaben IT stehen aber auch für Informationstechnologie, welche die Methodik der Lokalisierung andeutet. Konkret geht es darum, moderne Audio-, Bild- und Videoanalyse auf die verschiedenen Anwendungsszenarien anzuwenden. Letztere sind, dieser InnoProfileTransfer-Initiative entsprechend, aus dem Arbeitsumfeld der vier Stifterfirmen entliehen. So untersuchen die Arbeitsbereiche AB1 und AB3 die Lokalisierung und Klassifizierung von Personen und Objekten im Raum mittels Video und Audio. Diese Themen sind im Interesse des Stifters Intenta GmbH, zu deren Geschäftsmodell die Entwicklung und Vertrieb smarter Stereo-Sensoren aber auch automobiler Software zählt. Die 3D-Micromac AG entwickelt Großanlagen im Maschinenbau für die Halbleiter- aber auch Optikbranche. In deren Interesse steht entsprechend die bildbasierte Lokalisierung und Interpretation von Herstellungsprozessen, inklusive der darin involvierten Werkstücke wie z. B. zur Qualitätssicherung. Das zu erforschen ist Aufgabe des AB4. Die 3D-Insight GmbH hingegen entwickelt sogenannte *Powerwalls*, große Projektionsflächen, die von einem Verbund (von meist 6) zum System gehörigen Videoprojektoren zusammengesetzt werden. Da die Videoprojektortechnologie sich mit Blick auf verbesserte Farbbrillianz, Leuchtkraft und Bildauflösung permanent weiterentwickelt, ist es wichtig, methodisch auf der Höhe der Zeit zu bleiben, wenn es um die Entwicklung von Kalibriersystemen geht, aber auch um die echtzeitfähige Verarbeitung der auszugebenden Signale zu gewährleisten. Im AB5 des Projekts sollen Methoden entwickelt werden, die Effekte wie Bildverzerrungen und Farbfehler automatisch ermitteln und durch intelligente Algorithmen kompensieren, sodass diese flexibel auf neue Videoprojektortechnologien übertragbar sind.

Die Problemtypen und Herausforderungen, die alle Arbeitsbereiche einen, sind Fragestellungen der (i) Performanz (Echtzeitanalyse, Hardware- und Plattformabhängigkeit) und der (ii) Genauigkeit (Evaluation nach computerwissenschaftlichen Kriterien mittels etablierter Metriken wie *Recall* und *Precision* anhand gegebener Testdaten). Hinzu kommen Fragestellungen der semantischen Verknüpfung, Datenfusion, Datenhandling und der Synchronisation verschiedener Sensoren oder Prozesse.

Die Lokalisierung selbst lässt sich somit unterteilen in:

- Lokalisierung des Mediums an sich
- im Medium
- in der Welt

Ziel der beantragten Nachwuchsforschergruppe ist der Aufbau und/oder die Weiterentwicklung technischer Frameworks zur Bearbeitung dieser Lokalisierungsstrategien.

1.2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Zu den Vorarbeiten zählten die ebenfalls im Rahmen der InnoProfile-Initiative geförderten Projekte sachsMedia und Validax, die von der Professur Medieninformatik der TU Chemnitz eingeworben und erfolgreich durchgeführt wurden [24, 23]. Deren Schwerpunkte waren die Detektion und Erkennung von Personen und Objekten in Videomaterial regionaler Fernsehstationen, mit dem Ziel die Archive automatisch zu verschlagworten und eine erleichterte Durchsuchbarkeit zu ermöglichen, wodurch auch ein gewisses kulturelles Erbe bewahrt wurde. In beide Projekte maßgeblich involviert war Marc Ritter, der u. a. die Kontakte zur Intenta GmbH etablierte. Herr Ritter übernahm die Stiftungsjuniorprofessur und das dazugehörige Begleitprojekt LocalizeIT im August 2014. Weitere Kontakte zu den anderen drei Stifterfirmen ergaben sich aus der Historie der Fakultät für Informatik.

Als Stifterfirmen für die Stiftungsjuniorprofessur wurden folgende Firmen gewonnen, die während der Projektlaufzeit im regelmäßigen Austausch mit den Wissenschaftlern ihres Arbeitsbereich und der Projektleitung standen. Studentische Praktika, sowie Forschungs- und Abschlussarbeiten wurden an die Unternehmen vermittelt und durchgeführt.

Die Stifterfirmen und ihre Interessensgebiete seien im Folgenden kurz zusammengefasst:

- Intenta GmbH - Smarte Sensorentwicklung (Stereo-Sensoren) und *Automotive Software* Entwicklung, Zählung von Personen, Größen/Altersabschätzung, Verhaltensanalyse von Personen, Objekterkennung
- 3D-Micromac: Maschinenbau, Entwicklung für Halbleitertechnologie, Nutzung von bildgebenden Verfahren (Mikroskopie, Videosensorik), Charakterisierung von Laserverarbeitenden Prozessen, Qualitätsbeurteilung/Sicherung von Qualität
- 3D-Insight GmbH - Powerwall-Entwicklung mit Videoprojektoren, Optimierung der Bildqualität, Synchronisation, erweiterte Realität
- IBS Software and Research GmbH - Softwareentwicklung allgemein, Metadatenmanagement

1.3 Planung und Ablauf des Vorhabens

Das Projekt ist in 5 Arbeitsbereiche (ABs) aufgeteilt, bei denen es folgende Assoziationen mit den Stifterfirmen gibt: AB1 und AB3 - Intenta GmbH, AB2 - IBS Laubusch Software und Research GmbH, AB4 - 3D-Micromac AG und AB5 - 3D Insight GmbH. Alle ABs starten gleichzeitig und sind weitgehend unabhängig durchführbar. In Jahr 4 und 5 gibt es Überlapp, z.B. bei der Audio-Video-Fusion (AB1, AB3), aber auch bei der Aktivitätserkennung in (AB1, AB5) und der bildbasierte Verhaltensanalyse in AB1 und AB2. Der Projektablaufplan in Form eines Gantt-Diagramms ist in Abbildung 1 zu sehen. Darin enthalten sind auch die wesentlichen Kurzüberschriften, der in Quartalen abzuarbeitenden Projektinhalte. Die Zuordnung der Ergebnisse zu den jeweiligen Quartalen erfolgte in den 5 Zwischenberichten, die dem Projektträger am Ende der Jahre 2014 - 2018 übergeben wurden. Hier wurde zugunsten einer übersichtlicheren themenbezogenen Gesamtdarstellung der Ergebnisse auf eine quartalsbezogene Berichtsform verzichtet.

1.4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Der technische Stand zur Antragstellung umfasste unter anderem klassische *Computer Vision*- und *Machine-Learning*-Ansätze, wie *Support Vector Maschinen* (SVM, [112]), Adaptive Resonanz Theorie (ART, [34, 35, 48]), *Clustering* [127], etc. Diese Ansätze waren durch das Aufkommen von *Deep Learning* in 2012/2013 zum Projektstart plötzlich überholt [76]. Es gab in diesem Jahr einen riesigen Technologiesprung, so dass sich dieser auch auf das Projekt ausgewirkt hat. Die Methoden, zusammengefasst unter *Deep Learning* [78], durchdringen seitdem sehr breit die Wissenschaft, über technologieorientierte Unternehmen bis hin zu Anwendungen des täglichen Lebens.

Deep Learning glänzte vor allem in einen Bereich zuerst, der Erkennung von einer großen Menge von Objekten in einer Unmenge an Bildern (Objekterkennung auf der ImageNet Datenbank, [76, 128]), wo es ab 2013 jede andere Technologie um Welten schlug. Das ermöglichte in den Folgejahren dann eine ganze Reihe von Anwendungen, wie Objekt-, Personenerkennung, Lokalisation von Objekten in Szenen, etc. [76, 19, 97, 78].

Gleichzeitig ermöglicht *Deep Learning*, dass bisherige Ansätze der Künstlichen Intelligenz, wie *Reinforcement Learning*, plötzlich auf reale Anwendungen angewendet werden konnten, so dass dies die Entwicklung anderer Domains ermöglichte, wie im Bereich Spiele (Go [113, 114]), Computerspiele (Atari Spiele [91], Dota 2 [25]), und Robotiksteuerung [50].

Während nach Stand des Projektbeginns insbesondere CPU-basierte Ansätze der Bild- und Objektdetektion und -klassifikation sowie -generierung die Landschaft von Anwendungen wie beispielsweise im Bereich der Lokalisation und semantischen Verknüpfung von

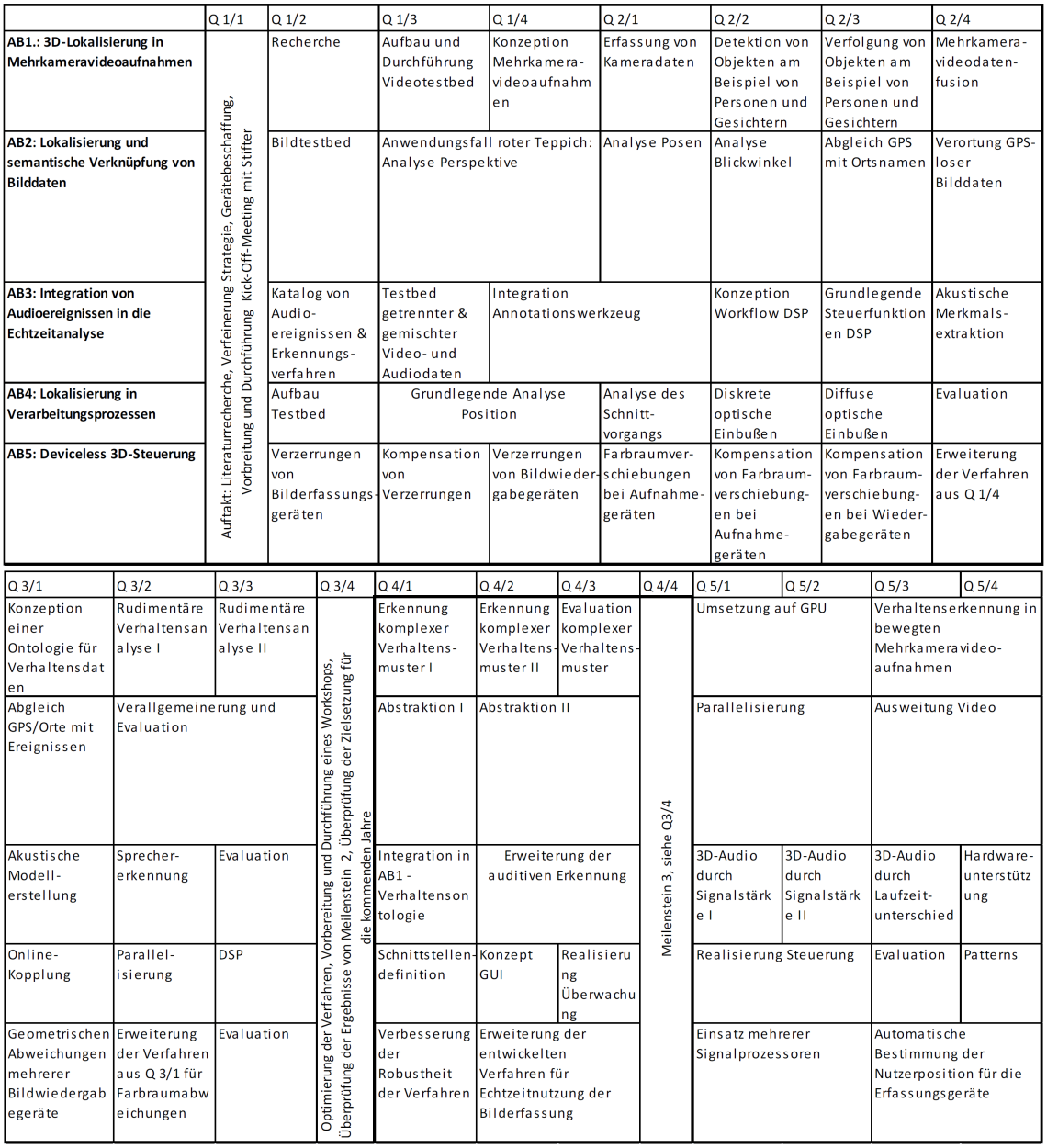


Abbildung 1: Überblick Projektplan.

Bilddaten, der Integration von Audioereignissen in die Echtzeitanalyse und Anwendungen kommerzieller, aber auch industrieller Anwendungsbereiche, wie beispielsweise zur Kontrolle von Herstellungsketten in der Halbleiterindustrie, prägten, ermöglichen die gegebenen Problemstellungen sowie die realisierten und konzipierten Verfahren die Integration komplexerer, insbesondere durch General Purpose Graphics Processing Units (GPGPU) beschleunigter Ansätze. So lässt sich nach aktuellem Stand der Technik ein Shift zu leistungsstärkeren Grafikkarten beobachten, deren Kommerzialisierung einen Einfluss auf die Verfügbarkeit im Kontext des wissenschaftlichen Rechnens nimmt. Dieser Technologie-

sprung in Form erhöhter Speichergrößen und einer erhöhten Parallelität durch eine Vielzahl paralleler Recheneinheiten, sowie die Verfügbarkeit durch Programmierschnittstellen wie beispielsweise CUDA wirken sich somit nicht nur auf die Möglichkeiten in ihrer Anwendung aus, sondern auch auf die aktuell zur Verfügung stehenden *Frameworks* im Bereich des *Machine Learnings*, die oftmals ihre eigenen Forschungsfragen und -communities bedienen. Diese umfassen unter anderem *Deep Learning Frameworks* und Programmbibliotheken wie TensorFlow, Keras, PyTorch und scikit-learn, Statistikapplikationen und Anwendungen des Data-Minings wie RapidMiner, Weka, und R, aber auch cloudbasierte Dienste wie Microsoft Azure und Google Cloud Platform sowie akademische Dienste, die beispielsweise über Anlaufstellen wie das Hochleistungsrechenzentrum der TU Dresden angeboten werden.

Daraus lässt sich ableiten, dass *Deep Learning* zwar auf den Objekterkennungsaufgaben wie ImageNet herausragende Ergebnisse liefert, aber dieser Durchbruch erst in andere Domänen zu transferieren ist.

Hierbei geht es um die Lösung bestimmter Forschungsfragen (themenspezifische Evaluationskampagnen), wo nach einem erfolgreichen Transfer durchaus $80 - 95$ % erreicht werden können. Allerdings ergeben sich auch eine ganze Reihe von Grenzen:

Es sind riesige Datenmengen nötig, was vor allem in den Fragestellungen für die lokale Industrie ein Problem darstellt. Zum Beispiel wenn es um die Erkennung von Fehlern in Werkstücken geht, sind dann doch einmal 1000 - 10000 defekte Werkstücke bzw. ihre Bildaufnahmen nötigt. Diese fallen oft im laufenden Betrieb nicht an, da Maschinenbauprozesse dann doch schon gut genug sind um nicht so eine hohe Ausschussrate zu haben und es mit riesigen Kosten verbunden wäre Werkstücke gezielt zu zerstören.

Ohne diese Datenmengen kommt es zu Problemen in *Deep Learning* wie *Overfitting* [78], was die Erkennungsraten mindert. Dem kann wiederum mit der künstlichen Synthese von Bildmaterial als Forschungsansatz begegnet werden. Oder als alternativer Ansatz, durch den Versuch bereits gelernte neuronale Netze auf das neue Problem zu transferieren (*Transfer Learning*, [118]), was aber, da das Bildmaterial sehr abweichend ist, auch eine Forschungsfrage für sich ist.

Gleichzeitig wird von der Industrie eine gewissen Robustheit und auch Erklärbarkeit der Klassifikationsentscheidungen gewünscht, da dem Kunden garantiert werden muss wie gut die Erkennung funktioniert, und auch unter welchen Bedingungen sie nicht funktioniert. Es hat sich in den letzten Jahren gezeigt, das Deep Netze stellenweise mit sehr abstrusen Bildmodifikationen zu Fehlentscheidungen neigen (*Adversarial Attacks*, [46, 77]), so dass dies auch eine Forschungsfrage ist.

Hauptsächlich müssen diese Punkte im Hinterkopf behalten werden, wenn es um den Transfer von *Deep Learning* in neue Industriedomains geht.

Gleichzeitig ist die aktuelle Forschung immer noch weit davon entfernt eine echte strong-KI zu verkörpern, denn es handelt sich meist um anwendungs-/themenspezifisch performante weak-KIs und keine Intelligenz nach Vorbild des menschlichen Gehirns.

Fachliteratur

In dem Vorhaben LocalizeIT wurden für die Bearbeitung der wissenschaftlichen Fragestellungen im wesentlichen auf Fachliteratur aus unterschiedlichen Quellen zugegriffen. Primär wurden Online-Recherche Werkzeuge wie google.scholar oder scopus.com verwendet. Daneben wurden aber auch neben der Universitätsbibliothek der TU Chemnitz und entsprechenden Online-Datenbanken (bspw. ACM Digital Library) Fachzeitschriften beschafft, anhand derer aktuelle Trends verfolgt werden konnten. Gleichzeitig war die Kommunikation mit den beteiligten Unternehmen unentbehrlich, da diese auf aktuelle Literatur für die einzelnen Forschungsgebiete und relevante Konferenzen hinweisen konnten. Die generell rege Teilnahme an entsprechenden Konferenzen rundete unser Vorgehen ab.

1.5 Zusammenarbeit mit anderen Stellen

Innerhalb der Region Chemnitz wurden Fragen des Projekts sowie mögliche Folgeanträge oder weitere Vorhaben mit themenverwandten Professuren diskutiert:

- Medieninformatik, TU Chemnitz
- Medieninformatik, HS Mittweida
- Professur Künstliche Intelligenz, TU Chemnitz
- Professur Technische Informatik, TU Chemnitz
- Professur Graphische Datenvisualisierung, TU Chemnitz
- Professur Schweißtechnik, TU Chemnitz

Besonders hervorzuheben ist die Zusammenarbeit mit den beiden Medieninformatik-Professuren mit denen seit 2018 jährlich Doktorandenworkshops unterschiedlicher Formate durchgeführt werden.

Mit den Stifterfirmen wurde sich regelmäßig ausgetauscht:

- Intenta GmbH
- 3D-Micromac AG
- 3D-Insight GmbH

- IBS Software and Research GmbH

Darüber hinaus entstanden Kontakte zu Firmen aus dem Bereich Automotive Software Engineering und Schweißtechnik:

- IAV GmbH
- Fusion Systems
- Harms und Wende QST GmbH
- EDC Chemnitz GmbH

Internationale Kooperationen:

Der Forschungsbereich Biochemie und Analyse von RNS-Interaktion mit Prof. Dr. Roland Sigel (Universität Zürich, Institut für Chemie, Metallo-RNA-Arbeitsgruppe) führte zu zahlreichen gemeinsamen Publikationen [31, 52, 30, 116].

- NIST/TRECVID organization committee
- Cornell Lab of Ornithology, Prof. Holger Klinck
- Universität Zürich, Institut für Chemie, Metallo-RNA Gruppe, Prof. Roland Sigel
- Florida Institute of Technology, Prof. Susanne Bahr

Nationale Kooperationen in Lebenswissenschaften:

Seit 2015 entwickelt die Stiftungsjuniorprofessur Media Computing gemeinsam mit der Augenklinik des Klinikums Chemnitz Projektideen im Bereich Augenmedizin und maschinellen Lernverfahren zur Therapieunterstützung. Es wurden drei Projekte erfolgreich eingeworben über das BMBF, die Sächsische Aufbaubank (SAB) und Novartis GmbH:

- Klinikum Chemnitz gGmbH, Augenklinik, Prof. Katrin Engelmann
- Sächsisches Makulazentrum (Augenkliniken in Dresden, Leipzig, Glauchau, Chemnitz, Aue), vertreten durch Prof. Katrin Engelmann
- Universitätsklinik Freiburg und Greifswald, Prof. Andreas Stahl
- Augen-OP-Zentrum Zschopau, Dr. Simo Murovski
- Biosaxony e.V., Andreas Hofmann
- Carus Consilium Dresden GmbH/Healthy Saxony e.V., Dr. Olaf Müller

- HS Mittweida, Professur Biophotonik, Prof. Richard Börner
- Smart Systems Hub Dresdens, Auge 4.0 Trail, Michael Kaiser
- Novartis GmbH, Mike Ross

2 Ergebnisse des Vorhabens und Vergleich mit der ursprünglichen Arbeits-, Zeit- und Ausgabenplanung

Die im Projektplan angestrebten Ziele konnten im Wesentlichen erreicht werden. Geringfügige Anpassungen wurden mit den Stiftern abgesprochen und werden in der Zusammenfassung der Arbeitsbereiche kurz erklärt und begründet. Der Überblick über die im Originalantrag festgehaltenen Kurzbezeichnungen der Arbeitspakete und deren zeitlicher Ablauf befindet sich in Abbildung 1. Hinweis: In Absprache mit dem Projektträger dürfen 306 statt 300 Personenmonate in der Projektlaufzeit abgerechnet werden, vorbehaltlich gegebener Finanzierung, aufgrund der personellen Fluktuationen, speziell beim Wechsel der Projektleiter nach der Berufung von Marc Ritter an die HS Mittweida nach 2-jähriger Laufzeit.

2.1 Verwendung der Zuwendung/Aufstellung der wichtigsten Positionen des zahlenmäßigen Nachweises

Die wichtigsten Positionen des zahlenmäßigen Nachweises sind in Tabelle 1 zusammengefasst.

Kategorie	abgerufen (€)	bewilligt (€)	Mehr/Minder (%)
0812 Beschäftigt	1.683.083,26	1.660.217,00	+1.38
0822 Wissenschaftliche Hilfskräfte	82,098.93	90,894.00 €	-9.68
0831 Gegenstände bis 400€	8,104.95	5,000.00 €	+62.10
0835 Verg. v. Aufträgen	20,000.00	20,000.00	0
0843 Sonstige allg. Verw.-ausg.	8,117.50	20,583.00	-60.56
0846 Dienstreisen	47,911.89	49,652.00	-3.5
0850 Gegenstände über 410€	276,588.69	280,156.00 €	-1.27
Gesamt	2,125,905.22	2,126,502.00	-0.03

Tabelle 1: Die wichtigsten Positionen des zahlenmäßigen Nachweises.

Hinweis: In der Position 0831 wurden im Originalantrag 10,000.00€ bewilligt. Die Position wurde in einem Umwidmungsantrag auf 5,000.00€ reduziert. Jedoch lag ein interner Buchungsfehler vor, der nicht mehr rückgängig gemacht werden konnte, so dass die 5,000.00€

versehentlich um mehr als 20% überschritten wurde. Eine Korrektur-Umwidmung wurde leider nicht mehr rechtzeitig beim Projektträger beantragt. Nimmt man aus Position 0843 5,000.00€ in Position 0831 rein, werden keine Einzelposten um mehr als 20% überschritten.

2.2 AB 1: 3D-Lokalisierung in Mehrkameravideoaufnahmen

Mehrkameraaufnahmen eines geschlossenen Raumes treten typischerweise in Überwachungsszenarien auf, seien diese sicherheitstechnisch, medizinisch oder wirtschaftlich motiviert. Hier gibt es bereits einige erfolgreiche Ansätze und Realisierungen auch bereits im kommerziellen Bereich. Herausforderungen bestehen aber nach wie vor darin Verdeckungen zu verarbeiten und robust auf unterschiedliche Signal-Rausch-Modalitäten (u.a. verursacht durch die Lichtsituation oder die Kameratechnik) einzugehen. Ein weiteres Problem liegt in der Übergabe von Objekten zu verschiedenen Kameras. Kann beispielsweise ein Raum nur durch den Einsatz mehrerer Kameras in seiner Gesamtheit erfasst werden, muss das verfolgte Objekt von einer Kamera zur nächsten übergeben werden. Im folgenden Unterkapitel werden zunächst die wichtigsten Aufgaben und Ziele aus dem Originalantrag zusammengefasst und deren Umsetzung im Projekt kurz skizziert. Im darauffolgenden Unterkapitel werden die jeweiligen Ergebnisse detailliert besprochen.

2.2.1 Aufgabenstellung und Zielsetzung

Im Projekt wurden Aufgaben quartalsweise formuliert und auf diese Weise auch in den Zwischenberichten abgehandelt. In der folgenden Auflistung sollen jedoch eher die übergeordneten Inhalte betrachtet und deren Lösung kurz angerissen werden:

1. Entwicklung von Ontologien zur Formalisierung von Verhalten:
   Es wurden rudimentäre Schemata entwickelt, die informationstechnologisch durch strukturierte Textdateien im json-Format abgebildet wurden. Diese orientieren sich am Standard, vorgegeben durch wissenschaftliche Evaluationskampagnen wie dem später noch detailliert vorgestellt werdenden TRECVID-Wettbewerb. Personen und Objekte, repräsentiert durch sogenanntes *Tags* werden mit Zeitstempeln (z.B. Video-Framenummer) versehen und Verhalten ebenfalls mit zeitgestempelten Tags umgesetzt.

2. Erfassung von Kameradaten und Verarbeitung von Mehrkameravideoaufnahmen:
   Es wurde ein Audio-Video-Labor errichtet, worin bestehende und neue Clustertechnik vernetzt wurde, um Mehrkameraaufnahmen zu ermöglichen. Zur Aufnahme wurde die API der smarten Stereosensoren intensiv getestet und für die Verwendung angepasst. Die Dokumentation von Ort und Orientierung im Raum basiert auf Kalibrierung der Kameras mittels Schachbrett. Dies erlaubt sowohl die Dokumentation als auch die spätere Modellierung und Simulation von Kameraperspektiven in der 3D-Simulationssoftware Blender.

3. Verfolgung von Objekten und Personen (im Raum):
   Hier wurden 2 Ansätze verfolgt: i. Aufnahme von Laborszenarien unter kontrollierten Bedingungen: Nachgestellt wurden Szenen mit vergessenen Gepäckstücken, Personen und Fahrrädern in einem Bus-Szenario, rote Teppich Ereignisse und Fahrräder im öffentlichen Raum. Zum Einsatz kamen klassische Verfahren (Vordergrund/Hintergrund-Trennung, Sparse Feature Detection), aber auch zunehmend CNN-basierte Methoden (*Convolutional Neural Network*). ii. Jährlich wurde an der Evaluationskampagne TRECVID teilgenommen, wo es um das Auffinden von Darstellern einer TV-Serien an verschiedenen Handlungsorten geht, aber auch im Aktivitätserkennung wie Telefonieren, Öffnen/Schließen, Ein-/aussteigen in/aus Auto, etwas ziehen/aufheben/abschließen. Dabei kamen CNN-basierte Methoden zum Einsatz. Ferner wurden die verschiedene Verfahren kombiniert und fusioniert. Nach dem Vorbild der Entwicklung eigener Bildverarbeitungs- und Metadatenhandlingsframeworks der Vorgängerprojekte „sachsMedia“ und „Validax“ ([24, 23]) wurde im Rahmen des Wettbewerbs eine neue Architektur entwickelt, die auf Webtechnologien und Verteilung setzt und moderne Bildverarbeitungsmethoden aus dem Bereich künstlicher Intelligenz über Virtualisierung effektiv einbindet.

4. Videodatenfusion:
   Einige der erstellten Szenarien wurden mit 3 gleichzeitigen Kameras-Streams aufgenommen und ausgewertet. Es konnte quantitativ gezeigt werden, wie die Fusion von mehreren Kameraperspektiven und mehrerer CNN-basierter Algorithmen die Zählung von Personen verbessert auf einen Fehler von 0.1 Personen pro Frame. Erwähnenswert ist, dass dabei auch klassische maschinelle Lernverfahren wie Entscheidungsbäume weiterhin Ihren Berechtigung haben und Ihren Mehrwert als Fusionsalgorithmus eindrucksvoll darstellten.

5. Einfache und komplexe menschliche Verhaltensanalyse und Evaluation:
   Neben der Entwicklung von Algorithmen war die Evaluation dieser stets im Fokus. Dazu wurde *Ground Truth* mit unterschiedlichen eigens entwickelten Labeling Tools erstellt. So entstanden einige Hunderttausend Labels mit Informationen über Ort von Personen und Objekten, Körpergelenkpunkte, Alter, Geschlecht, Hautfarbe, Bildähnlichkeit aber auch audiobasierte Annotationen. Diese dienten als Basis für die spätere Evaluation von Algorithmen zur Verhaltensanalyse im Rahmen des TRECVID-Wettbewerbs. Verhalten wurde dabei mehrheitlich durch die zeitliche Analyse von Körperpunkten bestimmt. Weitere *Ground Truth* wurde in den virtuellen Umgebung Blender und Unity automatisiert erstellt mit dem Ziel Grenzen bestehender Methoden wie OpenPose zu eruieren aber auch Modelle zu Verhaltensanalyse zu trainieren und auf die Realwelt (Labor-Szenarien, Trecvid-Videokorpus) anzuwenden.

6. Beschleunigung von Algorithmen und Nutzung von GPU:
   Da moderne Bildverarbeitungsalgorithmen (z.B. CNN-basiert) bereits die GPU nutzen (oft ca. 10x schneller als CPU) lag der Fokus der Untersuchung auf Beschleunigung durch Verteilung (bei vorhandener Hardware). Dazu wurden während eines Forschungsaufenthaltes am NIST (National Institute of Standards and Technology) Methoden der Virtualisierung evaluiert, um CNN-basierte Modelle einerseits zu verteilen, andererseits Modelle mit unterschiedlichen Abhängigkeiten gleichzeitig ausführen zu können. Dabei erwies sich Docker als die geeignetste Methode, die sich sogar auf portable/mobile Hardware wie Nvidia Jetson TX2 übertragen lässt. Es zeigte sich, dass damit in vielen Fällen Echtzeitanalyse von menschlichen Posen und damit Verhalten realisierbar ist.

7. Bewegte Kameras:
   Während klassische Methoden zur Objekterkennung sensibel auf veränderlichen Hintergrund und Licht reagieren, erwiesen sich CNN-basierte Methoden als deutlich robuster in Aufnahmen mit bewegter Kamera. Daraus ergab sich kein erhöhter Handlungsbedarf zur Weiterentwicklung von Algorithmen für bewegte Kameras. In Abstimmung mit den Stiftern wurde sich darauf konzentriert die immer neueren und besser performanten Architekturen neuronaler Faltungsnetze und vortrainierten Modelle eingehender zu untersuchen in Aufnahmen mit ruhenden Kameras.

2.2.2 Aufbau Labor: Audio-, Video, Bild

Kamera/Video-Ausrüstung

Drei Fotokameras („5Ds R“, Canon) und vier Videokameras („PXW FS7“, Sony) wurden zu Projektbeginn angeschafft. Diese dienen zum Erstellen von Bild- und Video-Testbeds, wie sie bspw. in den Bereich AB1 und AB2 notwendig sind. Dabei ist besonders die GPS-Fähigkeit hervorzuheben, die eine Teilaufgabe von AB2 darstellt, wo es um die Verschlagwortung von Bildern mit Ortsinformationen auf Basis von GPS-Koordinaten geht. Die Videokameras verfügen über bis zu 4K-Auflösung, jedoch können sowohl die Bildrate als auch die Auflösung angepasst werden, um auf bestimmte Testszenarien flexibel angepasst zu werden. Letzterem Ziel dient auch das beschaffte Stativmaterial, mittels dessen u.a. die Höhe und Winkel der Kameras justierbar sind. Die Fotokameras wurden bereits eingesetzt für die Erstellung des Testbeds „Roter Teppich“.

Audio-Ausrüstung

Die folgende Geräte wurden zu Projektbeginn zur Audioaufnahme beschafft:

- Audiointerface (Focusrite Scarlett 2i2)

- 2 Mess-Mikrofone (Behringer ECM-8000)
- 2 Mikrofonkabel
- 2 Mikrofonstative

Damit wurden erste Testdatensätze zur Audioklassifikation erstellt („10 Klassen *Ambient Assisted Living*“, siehe Kapitel 2.2.3).

Die Audio-Komponenten für das Media Computing (MC) Labor wurden gleich zu Beginn des Projekts beschafft. Die Abbildung 2 zeigt die Skizze des Audio-Video Labors. Das Labor besteht aus 2 Räumen. Im ersten Raum (B x T x H = 7.2 x 6 x 4 Meter) sind die akustischen und optischen Sensoren (smart sensor) installiert, auf einem Gerüst (B x T x H = 4 x 3.5 x 3.5 meter) an verschiedenen Positionen und Höhen mit dem Ziel der Lokalisierung und Verfolgung von Objekten. Dieser Raum wurde mit Schalldämmungsmaterialen eingerichtet. Im zweiten Raum befinden sich die Audiointerfaces im Rackschrank in einem gekühltem Serverraum.

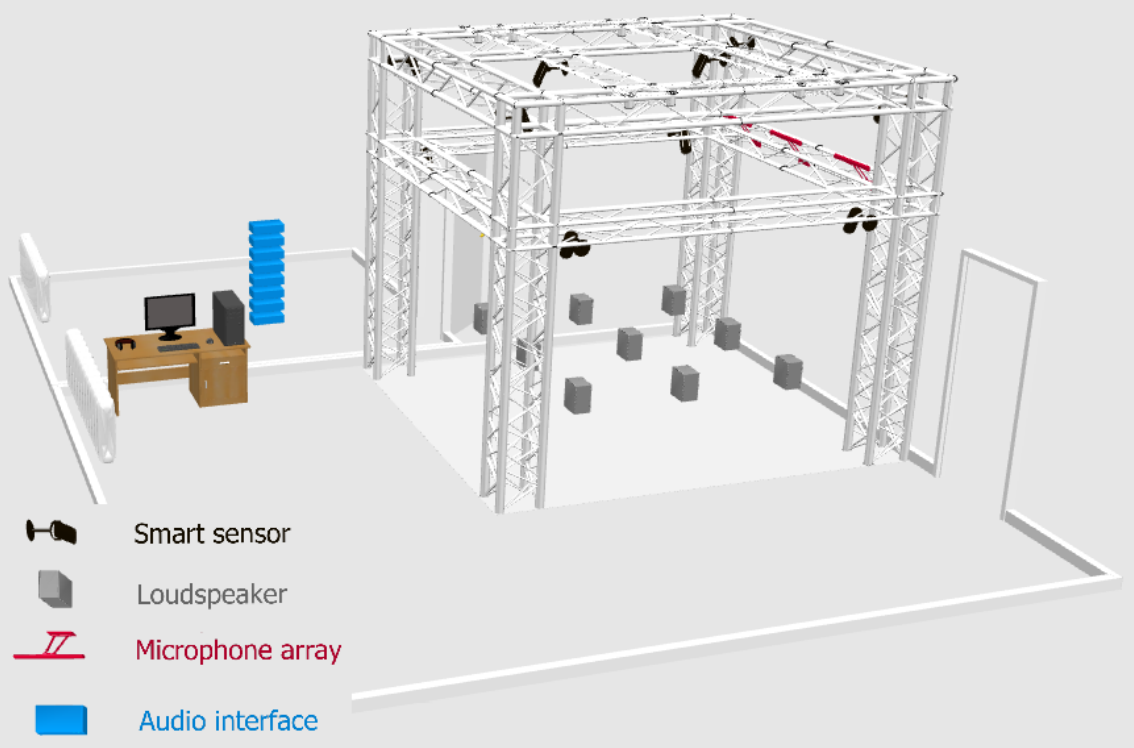


Abbildung 2: Skizze des MC-Labors mit der Darstellung der Positionen der akustischen und optischen Sensoren (smart sensors).

Die Abbildung 3 zeigt das Design der Audio-Komponenten. Verschiedene Mikrofone wurden zur Aufnahme der Audio-Daten benutzt. Die Anzahl der Mikrofone ist 64 und die Anzahl der Lautsprecher, die die Objekte simulieren, ist 16. Die Audiodaten werden beim Audiointerface vorverstärkt und digitalisiert. Jeweils 8 Mikrofone wurden zu einem Audio-

interface angeschlossen (es gibt 8 Audiointerfaces für die 64 Mikrofone). Die Audiointerfaces haben einen Alesis Digital Audio Tape (ADAT) Ausgang, der für die Übertragung der Audiodaten durch ein optisches Kabel zu einem Multi-Channel Audiointerface verbunden wird. Das Multi-Channel Audiointerface bietet Formatumwandlung von Multichannel Audio Digital Interface (MADI) [126] zu ADAT und umgekehrt. Die maximale Anzahl der Kanäle bei der Verwendung von MADI ist 64 (mit der Auflösung $24\ bit$ und bis $48\ kHz$ Abtastfrequenz). Die 8 ADAT optischen Eingänge im Multi-Channel Audiointerface werden auf MADI-Kanäle übertragen. Die Audiodaten können mit MADI-Kabel (koaxial oder optisch) mit einer Kabellänge von mehr als 100 m bis zur MADICard im Computer übertragen werden. Die Software *Steinberg Cubase Pro 8.5 EDU* [117] wird zur Aufnahme von Audiodaten der bis zu 64 Mikrofone in 64 einzelnen Kanälen verwendet und zur Generierung der Audiodaten vom Computer durch MADI-Kabel zu den Lautsprechern, wobei die MADI-Kanäle auf 8 ADAT optische Ausgänge übertragen werden und danach zu den Lautsprechern. Man benötigt entsprechend nur zwei ADAT optische Ausgänge für 16 Lautsprecher. Ein Audio-Master-Clock-System ist wichtig, um jedes Audiointerface als Slave mit der zentralen Master-Clock-Einheit zu verbinden (d.h. jedes Audiointerface bekommt den gleichen Takt vom Audio-Master-Clock-System). Die Bayonet Neill–Concelman (BNC) Kabel wurden zur Übertragung der Takte vom Audio-Master-Clock-System zu den Audiointerfaces verwendet. Verschiedene Arten von Mikrofonen - einschließlich kleine (Lavalier) und große (Mess-) Mikrofone - werden in jedem Mikrofon-Array getestet, um die Auswirkungen auf die Lokalisierungsergebnisse zu untersuchen. Darüber hinaus werden unterschiedliche Geometrien von Mikrofonarrays mit unterschiedlicher Anzahl von Mikrofonen untersucht.

Das Audio-System enthält die folgenden Komponenten:

- Mikrofone: Drei Mikrofonarrays wurden verwendet. In jedem Mikrofonarray wurden Mikrofone gleicher Art benutzt. Zusätzliche Mikrofone wurden zur Aufnahme von Sprach- und Musik-Daten benutzt. Die Tabelle 2 zeigt die technischen Daten der Mikrofone, die beim Mikrofonarray verwendet wurden.
  - Nowsonic Calibration Messmikrofon: 16 Nowsonic [8] Mikrofone wurden im ersten Mikrofonarray verwendet.
  - MXL 840 Mikrofon: Die 16 Mikrofone von MXL 840 [7] wurden im zweiten Mikrofonarray benutzt.
  - JustIn JM-714 Clipmic Mikrofon: Das dritte Mikrofonarray besteht aus 24 Mikrofone von JustIn JM-714 [6].
  - Audio Technica AT4040 Mikrofon: 2 Audio Technica AT4040 [5] Mikrofone wurden zur Aufnahme von Musik verwendet.
  - Rode NT5-MP Mikrofon: 2 Mikrofone von Rode NT5-MP [9] wurden benutzt.

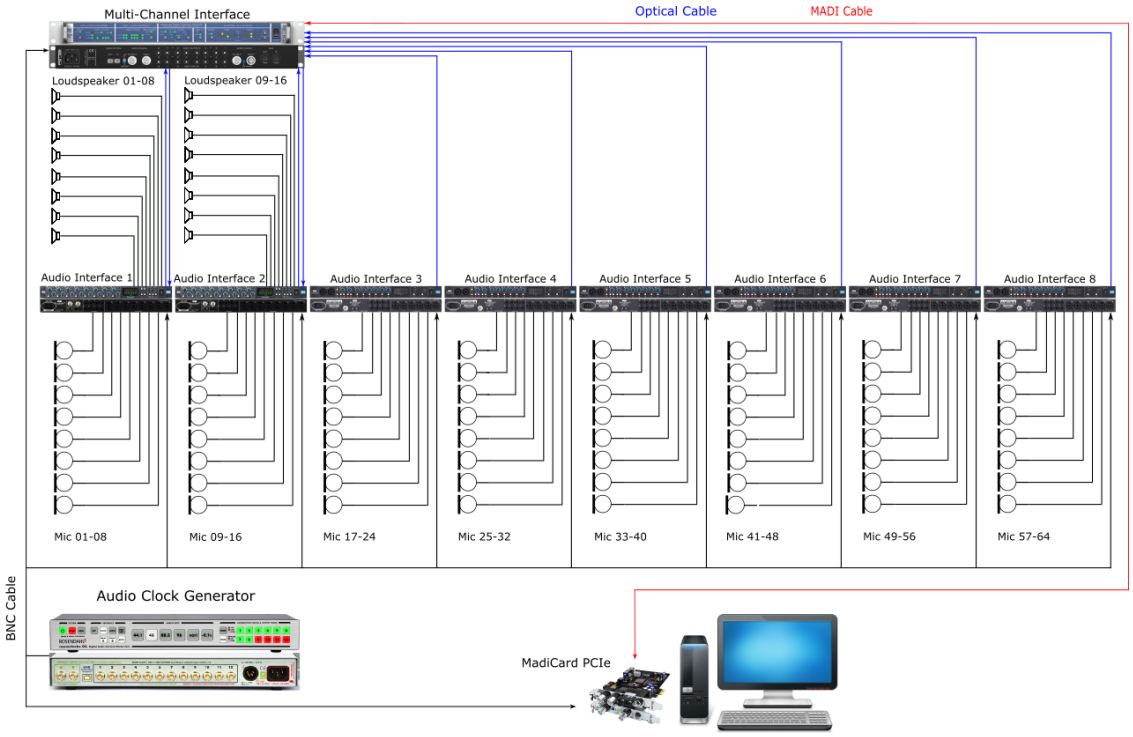


Abbildung 3: Schematische Darstellung von Audiokomponenten im MC-Labor.

- BLX288/PG5 Doppeldrahtlossystem [16] mit 2 handheld Mikrofonen.
- BLX188/CVL Dual Channel Lavalier Wireless System [15] mit 2 CVL Lavalier Mikrofone.

	Nowsonic	MXL 840	JustIn JM-714
Bauform	Elektret-Kondensator	Druckgradient-Elektret-Kondensator	Elektret-Kondensator
Durchmesser (mm)	6	22	< 5
Richtcharakteristik	Kugel	Niere	Kugel
Frequenzbereich (Hz)	20–20000	30–20000	50–16000
Ausgangsimpedanz (Ohm)	200	110	2000
Phantomspeisung (Volt)	12 – 52	48 ± 4	9 - 52
Abmessungen (mm)	21 x 200	22 x 134	
Gewicht (g)	140	280	
Preis (EURO)	55	50	40 (mit Zubehör)

Tabelle 2: Spezifikationen der Mikrofone bei der Mikrofonarrays

- Lautsprecher: 16 Lautsprecher wurden zur Simulation von Objekten verwendet. Die Tabelle 3 zeigt die Eigenschaften der Lautsprecher.
  - GENELEC: Die folgende Lautsprecher von Genelec [3] wurden verwendet.

    * Genelec 8010 AP
    * Genelec 8020 CPM
    * Genelec 8030 BPM
  - TANNOY: Die Lautsprecher von Tannoy [17] wurden verwendet.
    * TANNOY Reveal 402
    * TANNOY Reveal 502

	Genelec 8010 AP	Genelec 8020 CPM	Genelec 8030 BPM	TANNOY Reveal 402	TANNOY Reveal 502
Leistung (Watt)	25	20	40	50	75
Schalldruckpegel (dB)	96	105	108	101	108
Übertragungsbereich (Hz)	74 - 20000	66 - 20000	58 - 20000	56 - 43000	49 - 43000
Abmaße (HxBxT in mm)	195 x 121 x 116	226 x 151 x 142	299 X 189 X 178	240 x 147 x 212	300 x 184 x 238
Gewicht (kg)	1,5	3,7	5,6	5,2	7,2
Anzahl der verwendeten Lautsprecher	2	8	2	2	2

Tabelle 3: Eigenschaften der Lautsprecher

- Audiointerface:
  - RME ADI-648 [11]
  - Focusrite Octopre MK II Dynamic [2]
  - Focusrite Octopre MK II [1]
- Sound Card: RME HDSPe MadiCard PCIe [4]
- Audio-Master-Clock-System: ROSENDAHL NanoClocks GL [13]
- PreSonus Kopfhörerverstärker [10] und 2 Kopfhörer von Beyerdynamic
- Roland A-88 MIDI Controller-Keyboard [12]
- Yamaha DGX-660 WH Keyboard [14]

Die Audiodaten werden mit der Software (Steinberg Cubase Pro 8.5 EDU) aufgenommen (die Abtastfrequenz ist 48 kHz und die Auflösung ist 24 bit). Die Audiodaten von einem Mikrofon werden in einem Mono-Kanal gespeichert, d.h. 64 Audio-Dateien für 64 Mikrofone. Die Steuerung zum Abspielen der Audiodaten über die Lautsprecher wurde mit der Software (RME TotalMix) durchgeführt. Abbildung 17 zeigt den aktuellen Stand des Labors.

In Beitrag [40] wird die Realisierung des audiovisuellen Labors der wissenschaftlichen Community im Rahmen eines Fachbeitrags auf der ESSV (Elektronische Sprachsignalverarbeitung) 2018 vorgestellt. Darin geht es um dessen Nutzung zur Detektion, Lokalisierung, Klassifizierung und Verfolgung von Objekten in Innenräumen mit Hilfe von sowohl visuellen als auch akustischen Informationen. Mit den 10 intelligenten Stereosensoren (Typ S2000,

Abbildung 4: Positionen der Mikrofonarrays und Lautsprecher im audiovisuellen Labor.

Intenta GmbH) können Objekten und Personen erkannt und getrackt werden. Für die Audiosignalverarbeitung stehen insgesamt 64 Mikrofone und 16 Lautsprecher zur Verfügung. Die drei Mikrofon-Arrays bestehen aus drei verschiedenen Arten von Mikrofonen. Die Kamerasensoren und Mikrofone können in verschiedenen Positionen, Richtungen und Höhen befestigt werden. Die Lautsprecher können frei bewegt werden innerhalb der Tracking-Fläche mit Hilfe von Standard-Monitorständern. In einem separaten klimatisierten Serverraum werden alle Audiosignale vorverstärkt mit Hilfe von industrietauglicher AD/DA-Standard-Audio-Hardware zur Rackmontage. Ein Server-Cluster und Arbeitsstationen mit High-End Nvidia P6000-Grafikkarten bieten die Rechenleistung für die Analyse der Laboraufnahmen. Software, die für die Anwender im Labor zur Verfügung steht, umfasst kommerzielle Produkte, wie Steinberg Cubase 8.5, sowie selbst entwickelte Lösungen, wie z.B. ein Audio- & Video-Lokalisierungs- und Annotationswerkzeug [103]. Die Position akustischer Quellen (z.B. Lautsprecher) und Detektoren (Mikrofone) wurde in einer Blenderanimation in Form eines 3D-Labor-Modells dokumentiert (siehe Abbildung 29). Interessierte Personen haben auch die Möglichkeit, die zusätzliche Peripherie des Labors zu nutzen, z.B. ein Yamaha DGX-660 Keyboard, weitere Arten von mobilen Funkmikrofonen, HTC Wive VR-

Brille und drahtlose Smart Home Sensorsonden für die Bewegungserfassung. Die aktive Tracking-Fläche umfasst insgesamt 82 Sensoren.

Zu Projektbeginn wurden drei Fotokameras („5Ds R“, Canon) und vier Videokameras („PXW FS7“, Sony) angeschafft. Diese dienten zum Erstellen von Bild- und Video-Testbeds, wie sie bspw. in den Bereichen AB1 und AB2 notwendig sind. Dabei ist besonders die GPS-Fähigkeit hervorzuheben, die eine Teilaufgabe dieses Arbeitsbereiches darstellt. Die Videokameras verfügen über bis zu 4K-Auflösung, jedoch können sowohl die Bildrate als auch die Auflösung angepasst werden, um auf bestimmte Testszenarien flexibel angepasst zu werden. Letzterem Ziel dient auch das beschaffte Stativmaterial, mittels dessen u.a. die Höhe und Winkel der Kameras justierbar sind. Die Fotokameras wurden bereits eingesetzt und für die Erstellung des Testbeds „Roter Teppich“ genutzt.

2.2.3 Aufbau eigener Testsets: Audio-, Bild- und Videos

Im Projektverlauf entstanden zahlreiche eigene Datensätze, die im Folgenden gemeinsam mit anderen genutzten externen Datensätzen, kurz zusammengefasst werden.

- TV Studio/Bild/Video: Roter Teppich Szenario - Das Bildset wurde im Fernsehstudio der Professur Medieninformatik aufgenommen. Zur Selektion wurde die Qualität der Bilder hinsichtlich Beleuchtung, Rauschen und Sättigung mittels VIQET bestimmt [49]. Es beinhaltet über 3.500 Bilder und 6 Videos. Variiert wurden die Kamerahöhe und Position sowie die Posen und Blickwinkel von 11 verschiedenen Personen. Zur Nutzung des Datensatzes und der damit erzielten Ergebnissen, siehe Kapitel 2.3.2.
- TV Studio/Video: Autorennbahn - Das Szenario zeigt Autos auf einer Anki Overdrive Rennbahn. Es entstanden 16 Videos im mkv-Format mit 1080p Auflösung und einer Bildwiederholrate von 50fps. Die Videos dienten als Modellsystem zur Unterscheidung und zum Tracking von Fahrzeugen und zur Synchronisierung mehrerer Eingangsbilder. Es wurden auch 4k Videos erstellt mit fortlaufenden Aufnahmen aus 3-4 Blickwinkeln verschiedener Fahrszenarien mit Leerlauf und „noise“ durch Umsetzen der Autos per Hand zwischendurch.
- Im Kapitel 2.3.2 wird ein Datenset vorgestellt, das 1296 Bilder einer Person mit Kopf und Oberkörper zeigt, wobei in jedem Bild die 3D-Orientierung des Kopfes relativ zur Kamera auf wenige Grad genau bestimmt ist. Es dient zu Evaluation von Kopfposen-Algorithmen, aber auch als Exempel wie durch Annotation von Körperpunkten in 3D-Modellen alle Bilder, aufgenommen durch die 36 Kameras des Body-Scanners automatisch Labels bekommen, also die 2D-Koordinaten der Körperpunkte.

- Labor/Video: Bus Szenario - Die Aufnahmen wurden mit einer bzw. 3 Kameras durchgeführt, siehe Abbildung 19. Die Szene beinhaltet einen Sitz für den Busfahrer sowie 11 Sitzplätze für die Fahrgäste. Von den einzelnen Kameras wird in diesem Szenario meist nicht das gesamte Verkehrsmittel erfasst und es können Verdeckungen einzelner Personen auftreten. Die zur Analyse erhobenen Rohdaten bestehen aus 2 Videoaufnahmen, in denen typische Fahrgastsituationen nachgestellt wurden. Die erste Aufnahme erfolgte mit Kamera 3 und beinhaltet 3728 Videoframes. Die zweite Aufnahme wurde mit allen 3 Kameras gleichzeitig ausgeführt und beinhaltet bis zu 1780 Videoframes pro Kamera.

- Labor/Video: Das zuvor genannte Bus-Szenario wurde erweitert um kurze Vogelperspektivvideosequenzen mit Fahrrädern, jeweils ca. 200 Videos im Labor wie in Abbildung 17c zu sehen oder im Innenhof des Universitätsteils Straße der Nationen, siehe Abbildung 18. Während die Außenaufnahmen stehende Fahrräder zeigen, sind die Innenaufnahmen so gestaltet, dass kontinuierlich Fahrräder hinzukommen und unterschiedlich positioniert werden. Dabei werden sie auch von den schiebenden Personen partiell verdeckt.

- Labor/Video: eGate - Hier erfolgte die Aufnahme von Personen beim Passieren eines Gates aus senkrechter Vogelperspektive. Diese ließen unter kontrollierten Bedingungen kleinere und größere Testobjekte fallen wie z.B. Kreditkarten, Ausweise und Rucksäcke. Es entstanden 16 Sequenzen mit je 1300 - 2700 Bildern.

- Labor/Audio: 10 Klassen *Ambient Assisted Living* - Der Datensatz enthält 103 Minuten Laboraufnahmen im *.wav-Format aus den folgenden Klassen: help (175), scream (129), whimper (176), crying (192), quiet (45), strikes (475), vandalism (78), downfall of plate (84), dislocate furniture (126), chair movement (132). In Klammern sind jeweils die Anzahl der Beispiele je Klasse. Die Aufnahmen wurden mit 44.1 kHz (downsampled: 16 kHz) und 16 bit Auflösung erstellt. Die Erstellung der Aufnahmen erfolgte im TV-Studio der Medieninformatik der TU Chemnitz, sowie bei unseren Projektpartner „Intenta GmbH“. 2 Mess-Mikrofone wurden bei der Audioaufnahme verwendet. Das Alter der Personen, die bei der Audioaufnahme teilgenommen haben, lag zwischen 25 und 82 Jahre. 58 Personen nahmen an der Aufnahme der Audioereignisse, die von Menschen produziert werden (Hilfe, Schreien, Wimmern und Weinen) teil. 103 Minuten wurden insgesamt für die 10 definierten Ereignisse aufgenommen. Die Anzahl der Audiodateien nach der Annotation mit dem „Folker“ Tool umfasst 1.612 Dateien.

- Labor/Audio: 16 Statische Lautsprecher-Aufnahmen mit 56 Mikrofonen - Das Setting wurde, wie in Abbildung 52 zu sehen, mit 16 statischen Schallquellen und 3 Mikro-

fonarrays an der Wand mit je 16, 16 und 24 Mikrofonen in rechteckiger, planarer Anordnung und konstantem Abstand zwischen den Mikrofonen erstellt. Aufgenommen wurde ein etwa 10-sekündiges Radiojingle, bestehend aus Stille, Ansage und Jingle. Zur Nutzung des Datensatzes und der damit erzielten Ergebnissen, siehe Kapitel 2.4.5.

- Labor/Audio: Saugroboter beim Fahren durch das Labor - In diesem Videosetting wurde ein Saugroboter von oben aufgenommen und gleichzeitig das Audio-Signal mit den 3 zuvor genannten Mikrofon-Arrays aufgenommen. Der Roboter bewegt sich dabei eher zufällig, bzw. auf leicht gekrümmten Bahnen durch den Raum. Der Datensatz diente der Evaluation laufzeitunterschiedbasierter Audiolokalisation. Ergebnisse sind in Kapitel 2.4.5 zu finden.

- Virtuell/Blender: Media Computing Labor in verschiedenen Szenarien - Mit der 3D-Software Blender wurden Szenarien erstellt, welche die 3D-Positionen der verwendeten Sensorik (Smartsensoren und Mikrofonarrays) im Labor dokumentiert und gleichzeitig die wesentlichen Randbedingungen der Laborumgebung wie Wände und Türen nachvollziehbar dokumentiert. Ein Beispiel ist in Abbildung 29 zu sehen.

- Virtuell: 8 verschiedene menschliche Avatare bei verschieden Aktivitäten - In Kooperation mit Prof. Ritter, dem Leiter des Projekts während der ersten beiden Projektjahre, wurde gemeinsam ein Blender-Datenset im Rahmen der Lehrveranstaltung Lernfeld Wissenschaft und Wirtschaft an der Hochschule Mittweida erstellt. Dabei wurden mit MakeHuman Personen (Avatare bzw. Humanoide) erstellt und animiert mit dem Fernziel diese zur Evaluation in der Verhaltensanalyse einzusetzen. Die 8 entstandenen Avatare unterschiedlicher Größe, Geschlecht und Hautfarbe ist in Abbildung 5 zu sehen.

- Virtuell: Person mit Fahrrad aus Vogelperspektive mit 6 Videokameras - Das erstellte Blender-Szenario zeigt eine texturlose Ebene als Boden. Darauf führt eine weibliche Humanoidin mit 163 Knochen, blauer Kleidung und braunen Haaren Vorwärtsbewegung von 234 Schritten aus und führt ein leicht gefärbtes graues Fahrrad mit sich. Die Lichtabstimmung erzeugt nur wenige Schatten, wie an einem sonnigen Tag. Neun simulierte visuelle Stereosensoren sind in einer Linie, die orthogonal über der Mitte des Pfades zentriert ist. In den erzeugten 15 Videos ändert sich der Abstand zwischen den Sensoren zwischen 10cm, 30cm und 50cm und deren Winkel von -40° bis +40° in 20°-Schritten. Ein Beispiel ist in Abbildung 6 zu sehen. Damit wurden Ergebnisse zur Sensorfusion erzielt, die in Kapitel 2.2.9 näher erläutert werden.

- Virtuell: Businnenraum mit wenigen Personen - Eine synthetische Nachbildung eines Linienbusses mit Fahrer und Passagier wurde erstellt und ist beispielhaft in Abbil-

dung 26 zu sehen. Damit wurden Ergebnisse zur Personenerkennung abhängig von der Perspektive und Umgebung erzielt, die in Kapitel 2.2.9 näher erläutert werden.

- Extern-Audio: Birdnet Challenge - Der Datensatz besteht aus bis zu 1.500 Klassen mit Vogelstimmen und ist Teil einer Evaluationskampagne zur Vogelstimmenklassifikation namens BirdCLEF. Wir haben daran teilgenommen und eigene Challenges bei BirdClef erstellt. Es diente als Modelldatensatz für die *ambient assisted living*-Klassifikation in AB3, wo uns zunächst ein weitgehend annotierter Datensatz fehlte. Ergebnisse werden in Kapitel 2.4 eingehend vorgestellt.

- Extern-Audio: Google Audio Dataset - Es enthält mehr als 600 Audio-Klassen. Es wurde verwendet, um den Klassifikator für *ambient assisted living*-Kontext und für Negativkontrollklassen im BirdNET-Klassifikator zu trainieren. Beide Algorithmen werden im Kapitel 2.4 vorgestellt.

- Extern-Audio: ESC-50 - Es enthält 50 Klassen mit fünf Untergruppen wie Tier- und Menschengeräusche, wobei jede Klasse aus 40 Beispielhörproben besteht. Neben anderen eigenen und externen Audioressourcen wurde es es ebenfalls für die *ambient assisted living*-Klassifikation benutzt.

- Extern-Bild/Video: TRECVID Challenge - In dieser Evaluationskampagne gibt es verschiedene Unterwettbewerbe. Aufgrund der Ähnlichkeit der Fragestellungen im AB1 des Projekts wurde regelmäßig an der sog. *Instance Search Task* (kurz: INS) und später auch der *Activity Evaluation Task* (kurz: ActEv) teilgenommen. Details dazu werden in Kapitel 2.2.7 vorgestellt.

- Extern-Bild: Bilddatenbank der Universitätskommunikation der TU Chemnitz - Dieser Datensatz enthält über 30.000 Bilder der Universitätskommunikation der TU Chemnitz und enthält typische für Presse- und Publikationszwecke geplante hochqualitative Aufnahmen wie Sportfeste, Immatrikulationsfeiern, usw. Teile davon wurden genutzt um die automatisierte Verschlagwortung von Bildern in AB2 zu evaluieren. In Kapitel 2.3 wird darauf näher eingegangen.

2.2.4 Nutzung vorhandener und Entwicklung eigener Annotationstools

Zur Erstellung von *Ground Truth* aber auch zum Trainieren von Algorithmen wurden Audio, Bild- und Videodaten annotiert. Es ergaben sich jedoch unterschiedliche Anforderungen, je nach Aufgabenstellung, so dass neben bestehenden auch eigene Annotationstools (weiter)entwickelt wurden.

Abbildung 5: 8 mit MakeHuman[1] erstellte Charaktere mit unterschiedlicher Hautfarbe, Geschlecht, Alter und Größe.

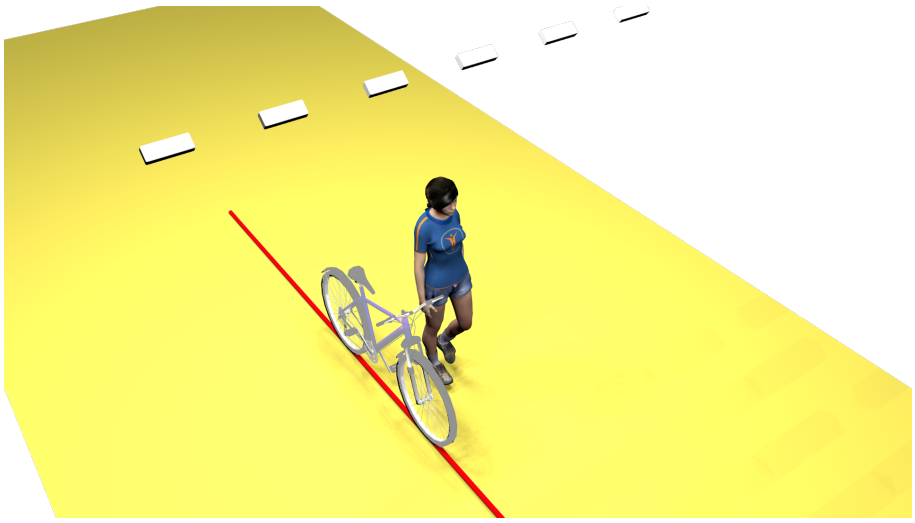


Abbildung 6: Mit dem Programm Blender erstelltes virtuelles Szenario: Eine Frau läuft mit Fahrrad unterhalb mehrerer Video-Sensoren. Der Abstand zwischen den Sensoren sowie deren Winkel wurden systematisch geändert. Für Details, siehe [86].

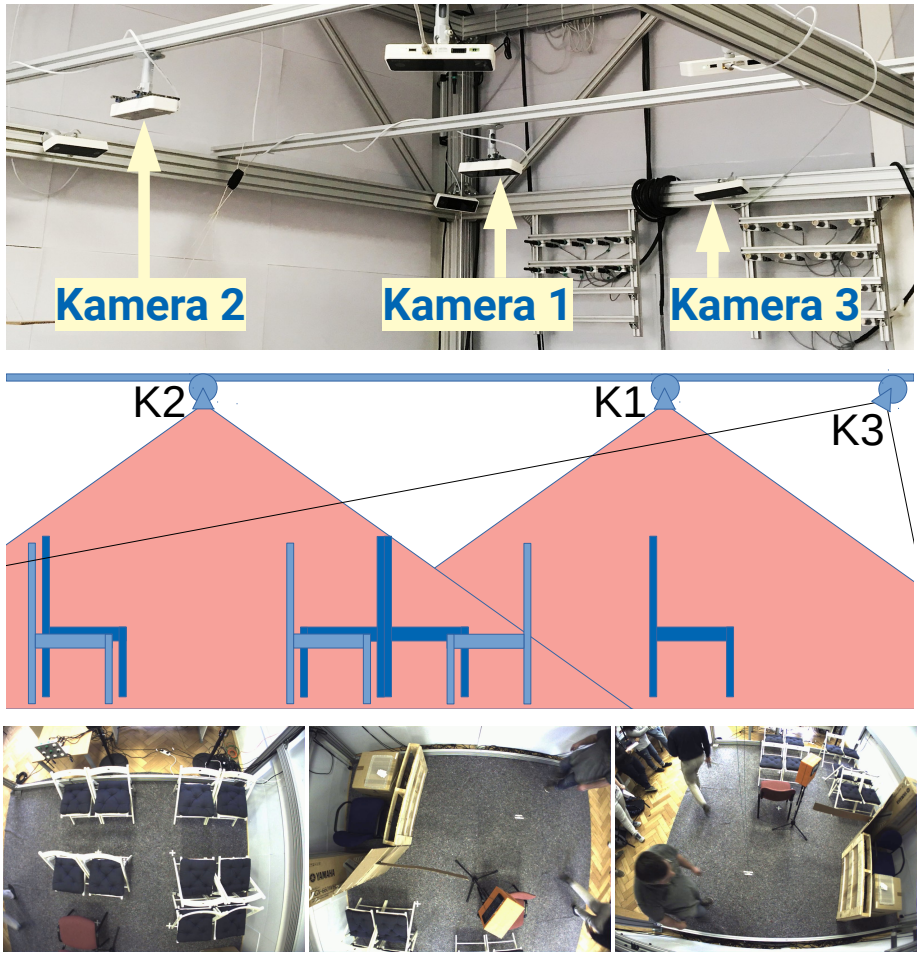


Abbildung 7: Laboraufbau für die Busszene mit 3 Kameras

- Extern: LabelMe - In Vorbereitung der Posen- und Blickwinkelanalyse wurde ein Großteil der Bilder des Datensatzes „Roter Teppich Szenario“ systematisch annotiert mit rechteckigen Zeichen-Boxen (engl. bounding box) unter Verwendung des Online-Tools „LabelMe“ [2]. Die Annotationen werden hier im XML-Format zur Verfügung gestellt und sind damit einfach in eine Datenbank übertragbar.

- Extern: Geosetter - Die Annotation GPS-basierter Daten erfolgte mit dem Freewareprogramm „Geosetter V. 3.4.16“. Dieses erlaubt die Visualisierung von in Bildern enthaltenen GPS-Koordinaten in Karten von Google und OpenStreetMaps, aber auch die Annotation von GPS-losen Bildern mittels nutzerdefinierter GPS-POIs (POI - Points of Interest). Darüber hinaus ist es ein leistungsstarkes Annotationstool mit immenser Funktionalität zur Metadatenanreicherung. Es benutzt im Hintergrund „ExifTool“ von Phil Harvey[3], welches es ermöglicht annotierte Bild-Metadaten direkt in Felder des EXIF- oder IPTC-Dateiheaders zu schreiben oder als *.xps Datei zu hinterlegen. Au-

[2] http://labelme.csail.mit.edu/Release3.0/
[3] http://www.sno.phy.queensu.ca/~phil/exiftool/

ßerdem erlaubt Geosetter das automatische Abrufen von Ortsinformationen wie Ländercode, Land, Bundesland, Stadt und Ort (Stadtteil). Es kam zur Metadatenanreicherung von Bildern mit Ortsdaten für den AB2 zum Einsatz. Näheres, siehe Kapitel 2.3.

- Extern: Folker Tool - (der FOLK EditoR) wurde für das Projekt „Forschungs- und Lehrkorpus Gesprochenes Deutsch“ (FOLK) am Institut für Deutsche Sprache entwickelt. Das Datenmodell für dieses Werkzeug stellt eine Adaption des EXMARaLDA-Systems dar[4]. Es erlaubt als Standalone-Programm die Segmentierung von Audiodateien, die anfänglich im AB3 mit FOLKER durchgeführt wurde innerhalb der Audioaufnahmen, die für den *Ambient Assisted Living* -Kontext erstellt wurden.

- 3D-Annoations-Tool für Body-scanner: Dieses Tool erlaubt im Body-Scanner-Setup der Professur Graphische Datenverwaltung der TU Chemnitz das entstandene 3D-Modell an 7 nutzerdefinierten Punkten zu labeln, hier z.B. Augen, Ohren, Nase und Mund. Im Hintergrund wurden daraus die Koordinaten in den 36 zugehörigen Kamerabildern bestimmt, basierend auf deren Kameramodell. Ferner ließen sich daraus auch die Winkel der Normale der Ebene aus Augen und Nasenspitze relativ zur Kameranormale bestimmen.

- Audioklassen-Label Tool für Videos: Dieses Tool ist eine sog. *Progressive Web App* und erlaubt das Labeln kurzer Videosequenzen mit Audio-Klassen. Es wurde zur Erstellung von *Ground Truth* im TRECVID-Wettbewerb eingesetzt.

- Extern: Vatic - Dieses Programm ist speziell zur Annotation von Videos nützlich, da es zwischen mehreren Bildern interpolieren kann und kontinuierliche Bewegungen von Objekten nicht Frame für Frame abgearbeitet werden müssen.

- Intern: Deep Learning Web Lab Annotator (DLWLA) - Das browserbasierte Programm erlaubt vor allem das kooperative Annotieren von Bildern mit einfachen geometrischen Primitiven wie Punkten, Polygonen, Rechtecken, Quadraten und Kreisen, aber auch binäre True/False und 3-7 Sterne Bewertungen. Der häufigste Einsatz war das Labeln von Personen und Posen z.B. für Personenerkennung Rahmen des TRECVID-Wettbewerbs. DLWLA wurde erweitert, um modellgetrieben zeiteffizient annotieren zu können, d.h. die vom Posenerkennungsverfahren Open Pose (Modell) gelieferten Körper(gelenk)punkte werden mit ihrem jeweiligen Label (z.B. linke Schulter hellgrün in Abbildung 8) bereits angezeigt und müssen nur noch korrigiert werden, falls Sie falsch lokalisiert wurden. Nicht erkannte Körperpunkte können neu hinzugefügt werden, während nicht sichtbare entfernt oder als nicht sichtbar annotiert werden können. Für das rote Teppich-Testset wurde ausschließlich die Person

[4] `https://exmaralda.org/de/folker-de/`

auf dem roten Teppich analysiert. Damit entstanden beispielsweise ca. 1000 Bilder mit bis 18.000 (*Ground Truth*) x,y-Koordinaten aus den Kategorien Füße, Knie, Hüften, Schultern, Ellbogen, Handgelenke, Hals, Augen, Nase, Mund und Ohren jeweils nach Körperseite unterteilt.

- Intern: Audio, Bild, Videoannotationsframework von Christian Roschke - Im Rahmen der Masterarbeit von Christian Roschke wurde ein webbasiertes Managementsystems entworfen und implementiert. Es dient zur Annotation von Audio-, Bild- und Videomaterial als Basis für die Entwicklung maschinell trainierter Algorithmen. Mittels Laravel und OctoberCMS wurde die konzeptionierte Architektur umgesetzt. Dabei kann serverseitig mit der Programmiersprache PHP und clientseitig mit JavaScript, HTML und CSS gearbeitet werden. Das System wurde im Rahmen der TRECVID-Wettbewerbe 2017 bis 2019 erweitert um eine Vielzahl verteilter Komponenten zur Entwicklung von Algorithmen der Personen-, Objekt- und Ortserkennung. Eine Zusammenstellung der grafischen Nutzeroberflächen findet sich in den Abbildungen 9, 10 und 11. Die in Abbildung 49 dargestellte Architektur basiert auf dem Client-Server-Prinzip. Auf der Client Seite wird dabei einem menschlichen Nutzer mittels UI (*User Interface*) und einer Maschine mittels API, die Interaktion mit dem Client-Controller ermöglicht. Dieser leitet Anfragen an die Schnittstelle zur Server Applikation, dem Communication-Handler, weiter. Der Communication-Handler ermöglicht das Senden und Empfangen von Nachrichten über AJAX, direkte Verbindungen, Server-Sent-Events (SSE) und Websockets

- Intern: Videobasiertes Audioklassen-Label-Tool - Ein als Progressive Web App entwickeltes Programm dient der manuellen Klassifikation von Audiodaten. Darin können ordnerbasiert Videos geladen und wiedergegeben werden. Eine importierte Liste von nutzerdefinierten Audio-Klassen dient der Zuordnung der Videosamples in diese Klassen. Es eignet sich besonders für kurze Videosequenzen, wo es nicht auf exakte Zeitstempel ankommt. Es wurde zur Erstellung von *Ground Truth* im TRECVID-Wettbewerb eingesetzt, wo sog. *shots* (filmische Szene zwischen 2 Schnittgrenzen) annotiert wurden. Die annotierten Daten dienten als Trainingsdaten für einen CNN-basierten Audioklassifikator, der z.B. weinen, essen oder schreien als relevante INS-Tasks identifizieren sollte.

2.2.5 Aufbau Cluster, Storage, GPU-Workstations und mobilen Recheneinheiten

Neben der Wartung und Einbindung des aus dem Vorgängerprojekt sachsMedia bestehenden Clusters, intern bezeichnet mit den Namen „Skinner“, wurde auch ein neuer Rechencluster eingerichtet. Skinner besaß 12 Knoten mit je 4 Kernen, die Hyperthreading unter-

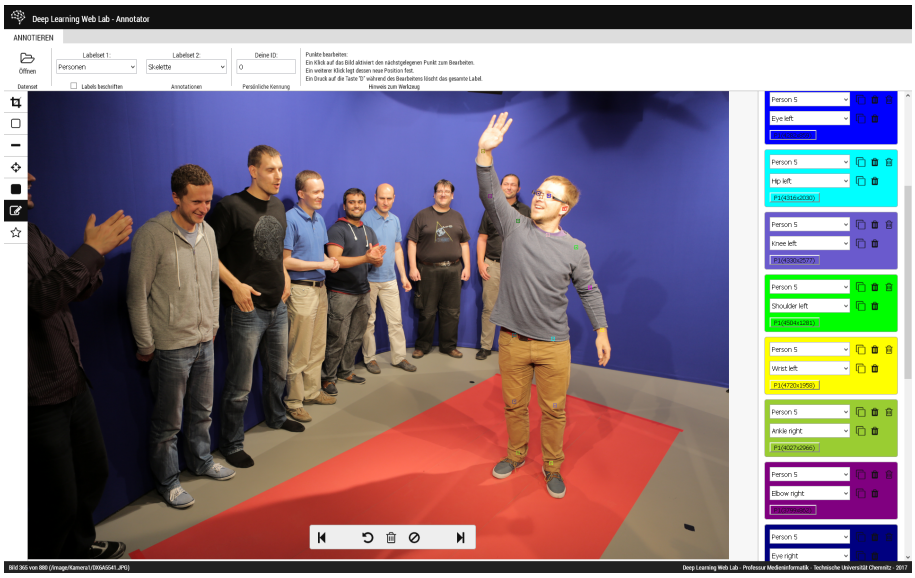


Abbildung 8: Screenshot eines webbasierten Annotationstools, das kollaboratives Arbeiten erlaubt hier am Beispiel farbig markierter Körper-Punkte, die von einem Verfahren wie [125] geliefert werden und durch Mausklick-Gesten zur Korrektur verschoben werden können.

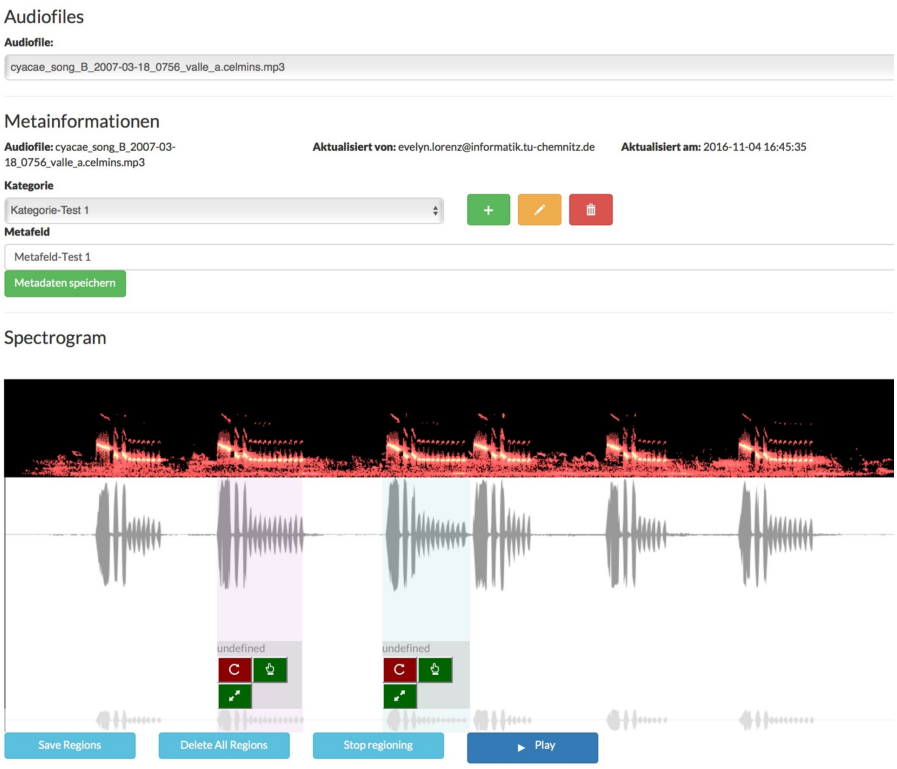


Abbildung 9: Grafische Nutzeroberfläche des Moduls zur Audiosegmentierung aus Christian Roschkes webbasierten Managementsystems zur Annotation audiovisueller Medien

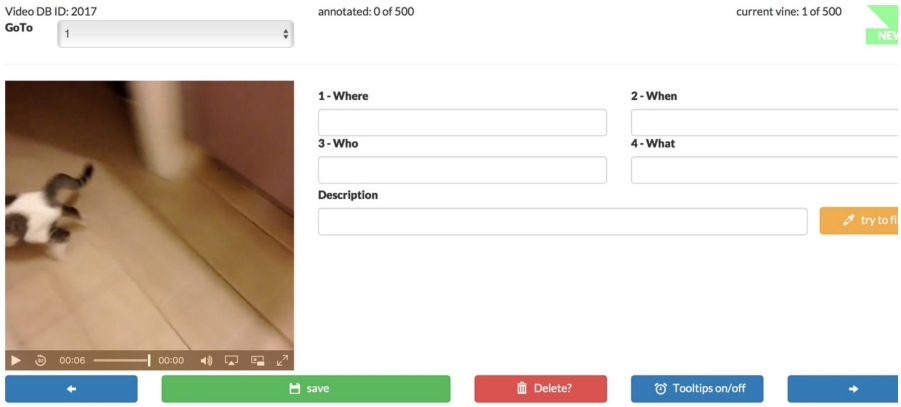


Abbildung 10: Grafische Nutzeroberfläche des Moduls zur Video-Metadatenanreicherung aus Christian Roschkes webbasierten Managementsystems zur Annotation audiovisueller Medien

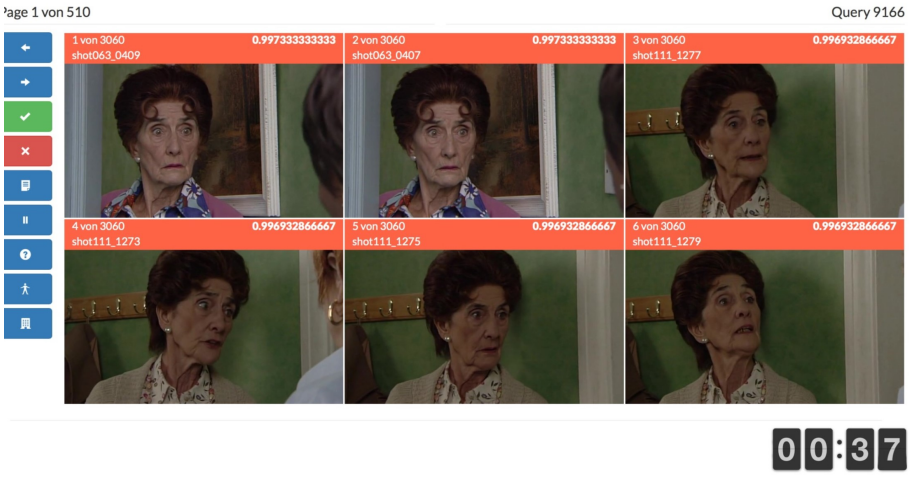


Abbildung 11: Grafische Nutzeroberfläche des Moduls zur Bild-Metadatenanreicherung aus Christian Roschkes webbasierten Managementsystems zur Annotation audiovisueller Medien: Der Nutzer wählt unter den 6 Bildern, das oder diejenigen aus die ein zuvor bestimmtes Kriterium erfüllen.

stützen. Insgesamt standen darauf 60TB Speicher zur Verfügung, genutzt in der Raid-Stufe 5 oder 6. Jeder Knoten bot zudem ca. 16GB RAM und je ein Windows Server und Linux. Damit wurden besonders in der Anfangszeit noch CPU-lastige Rechenaufgaben durchgeführt, z.B. in der Vordergrund-Hintergrund Trennung. Die enthaltenen Grafikkarten waren allerdings für die Anforderungen im Bereich *Deep Learning* nicht mehr ausreichend. Daher wurden neben den im Projekt vorgesehenen 2 GPU-Workstations mit P6000-Grafikkarten der Firma Nvidia weitere Consumer-Grafikkarten beschafft und in die Arbeitsplatzrechner der Projektmitarbeiter intergriert. Maßgebliche Erfolge ließen sich durch den Einsatz der Grafikkarten in der Anwendung CNN-basierender Algorithmen erzielen wie z.B. im Bereich Personenerkennungung, Posenanalyse und Ortserkennung. Ferner wurden eigene Algorithmen unter langfristiger Nutzung der GPU-Workstations entwickelt, etwa in der Erkennung von Fahrrädern (AB1), Vogelstimmen (AB3) und Audioklassen im *Ambient Assisted Living* Kontext (AB3).

In der zweiten Projekthälfte wurde ein neuer Rechencluster beschafft, der insgesamt 190TB Speicher zur Verfügung stellt und die folgenden Recheneinheiten enthält:

- 4 $\times$ Intel Xeon CPU E5-2695v4 @2,1GHz mit 18 Kernen und Hyperthreading, 64GB RAM
- 4 $\times$ Intel Xeon CPU E5-2695v4 @2,1GHz mit 18 Kernen und Hyperthreading, 128GB RAM

Betriebssystemunabhängiger Zugriff auf Speicher bzw. Daten für alle Projektmitarbeiter wurden über Samba-Protokoll (Netzlaufwerke) und per Nextcloud-Einrichtung gewährleistet. Letztere diente dem besseren Austausch mit Stiftern und Studierenden, aber auch der lokalen Synchronisierung und Backup spezifischer Daten. Genutzte Betriebssystem waren: Linux Debian, Ubuntu und Centos, Windows 7, 8, 10 und Server 2012r2. Der Speicherumfang von 190TB war notwendig, um eigens erzeugte Aufnahmen, aber auch externe wissenschaftliche Datensätze zu hosten, zu sichern und zu verteilen. Details zu den projektrelevanten Datensätzen sind in Kapitel 2.2.3 zusammengefasst. Unter den genannten Daten nahmen Videodaten erwartungsgemäß den größten Anteil ein. Diese entstanden maßgeblich in AB1, aber auch im AB4. Auf dem Cluster wurden zahlreiche Server eingerichtet, bspw. zur Netzwerk- und Rechnerverwaltung, Datenverteilung und -sicherung, Datenspeicher- und VM-Steuerung (VM - Virtuelle Maschinen) und zur Audiosteuerung. Ferner erlaubt die Rechenkapazität den Betrieb zahlreicher VMs zur Sensorsteuerung, Posenanalyse, Inhalts-, Gesichts- und Umgebungserkennung, Metadatenmanagement-Tool, Curvefit-Webanwendung, Annotationstools, Dokumentenverwaltung, Quellcode-Versionierung, Datenbanken, Content Managemangent System und für studentische Arbeiten.

Gleichzeitig wurde die Erprobung der Nutzbarkeit der vorhandenen Detektoren und Auswertungslösungen auf Embedded-Systemen weitergeführt und die zwei im Projekt beschafften *Nvidia Jetson TX2* benutzt. Mit dieser kompakten Bauform ist die Erprobung und Anwendung der entwickelten und angewandten Systeme auf Hardware möglich, welche der des S2000 des Kooperationspartners Intenta nahe kommt. Um der Flexibilität ähnlich normaler Rechner zu ermöglichen wurde auch hier der Einsatz der Virtualisierungslösung Docker erprobt und z.T. umgesetzt. Einzelne Hardwarebestandteile, wie z.B. die Onboardkamera, konnten hierfür nicht genutzt werden, wodurch die vollständige Vergleichbarkeit nur eingeschränkt möglich ist. Andere Komponenten wie die enthaltene GPU sind einsetzbar und erlauben den Einsatz der entwickelten und angewandten Systeme auf Videos mit bis zu 8 fps.

2.2.6 Objekt/Personenerkennung/tracking I: Klassischer Vordergrund-/Hintergrund-Trennung

In Absprache mit den Kooperationspartnern wurden Verfahren zur Vordergrund-Hintergrund-Trennung (Abbildung 12) detaillierter untersucht, um die Robustheit der bisherigen Verfahren gegenüber Rauschen oder Helligkeitsschwankungen zu verbessern und damit die Detektionsqualität nachfolgender Verfahren zu erhöhen. Speziell bei relativ kleinen oder schlechter differenzierbaren Objekten wie Kreditkarten, Ausweisen, Kaffetassen oder hellem Papier vor hellem Hintergrund (Abbildung 13), weisen die bisherigen Trennungsverfahren Defizite auf. Zur Anwendung kamen alle Verfahren der BGSLibrary aus OpenCV von Sobral *et al.* [115]. Damit wurde der in Kapitel 2.2.3 vorgestellte Datensatz „Labor/Video: eGate“ umfassend analysiert und evaluiert. Dafür wurde ein eigens entwickeltes Matlab-basiertes Evaluationsprogramm genutzt, dass es erlaubt in den erzeugten Grauskalenvideos schwellwertbasiert und unter Verwendung der *Ground Truth*-Bounding-Boxen die Erkennung und Verfolgung von Objekten und Personen zu evaluieren. Dabei wurden auch eigene Verfahren evaluiert, die zwar in der *Precision* mit bis zu 90% mithalten konnten mit den über 30 Algorithmen, jedoch nicht im *Recall*. Abschließend an diese Ergebnisse wurde gemeinsam mit den Stiftern beschlossen diese Art von Algorithmen nicht mehr eingehender zu untersuchen, sondern sich auf moderne, robusterer Verfahren der Objekt- und Personenerkennung zu fokussieren, nämlich den überwachten Lernverfahren basierend auf neuronalen Faltungsnetzen.

2.2.7 Objekt/Personenerkennung/tracking II: TrecVid Wettbewerb zur Evaluation von Algorithmen zu Objekt-, Personen-, Orts- und Verhaltenserkennung

Während der gesamten Projektlaufzeit nahm die Stiftungsprofessur Media Computing in Zusammenarbeit mit den Professuren Medieninformatik der TU Chemnitz (Prof. Eibl) und seit 2017 auch der Hochschule Mittweida (Prof. Ritter) an der international ausgerichteten wissenschaftlich renommierten Evaluationskampagne TREC Retrieval Evaluation on Videos (TRECVid) teil. Hier ist es den Nachwuchswissenschaftlern gelungen Spitzenplätze in Teilbereichen der Kategorie Instance Search zu erreichen durch die Kombination klassischer maschineller Lernverfahren mit modernsten Algorithmen wie *Convolutional Neural Network* (CNN). Die Ergebnisse wurden im November jeden Jahres auf dem TRECVid Workshop des National Institute for Standards and Technology (NIST), Gaithersburg, Maryland, USA, vorgestellt. Dabei wurden auch die bestehenden Kontakte an die Stiftungsjuniorprofessur, vertreten durch Robert Manthey und Stefan Kahl, weitergegeben, um die Nachhaltigkeit dieser Kooperation über 2016 hinaus nach dem Weggang von Prof. Ritter zu gewährleisten. Hervorzuheben ist noch, dass die IPT-Initiative in diesem Jahr erstmals aktiv einen Pilot-Task des TRECVid Workshops in der Kategorie „Video to Text Description“ (VTT) mitgestaltete. Der dazugehörige Artikel wurde über 100 mal zitiert und ist damit der meistzitierte Artikel des Projekts [41]. Im VTT-Task kam das webbasierte Managementsystems zur Annotation audiovisueller Medien von Christian Roschke zum Einsatz, ganz konkret das in Abbildung 10 dargestellte Tool mit dem einige Tausend Vines (6 Sekunden Videos) annotiert wurden.

Die o.g. Vernetzung ermöglichte auch einen direkten Forschungsaufenthalt beim NIST. Dieser beinhaltete die Erkundung der Möglichkeiten und Grenzen öffentlich zugänglicher und kosteneffizienter Virtualisierungslösungen wie VirtualBox, Xen, KVM und Docker, wobei auch die Nutzung von Consumer-GPU-Systemen wie den dort vorhandenen Nvidia Titan-X mit betrachtet und vor dem Hintergrund der Arbeitsaufgabe „Beschleunigung von Algorithmen und Nutzung von GPU“ auch als projektrelevant eingestuft wurde. Im Ergebnis wurde Docker als flexibelste und leistungsfähigste Lösung eruiert, welche den vollen Leistungsumfang allerdings nur mit dem Betriebssystem Linux und der Variante *Nvidia-Docker* erreicht.

Die vorhandenen Detektions- und Analysesysteme für Einzelbilder wurden, aufgrund ähnlicher Vorgehensweisen, im Aufgabenbereich *Instanz Search* (INS) angewandt. Demgegenüber kamen die Systeme für Bewegtbilder und Videos, wie z.B. Objekttracker, beim Aufgabenbereich *Activities in Extended Video* (ActEV) zum Einsatz [124]. Bei INS stellte sich die Aufgabe aus dem Beispielmaterial der britischen *BBC*-Fernsehserie *EastEnders* mit fast 500 h Videomaterial und dementsprechenden 42 Millionen Einzelbildern, 28 Hauptdarstellern und über 30 Handlungsorten, die vorgegebenen Personen an vorgegebenen Orten im gesamten Videodatensatz zu finden, wie in Abbildung 14 zu sehen.

Bei ActEV stellte sich hingegen die Aufgabe im VIRAT-Datensatz[5] alle Personen und Fahrzeuge zu finden, welche eine spezifizierte Aktivität ausführen, wie in Abbildung 15 zu sehen.

Durch diese Teilnahme war auch eine Einreichung und Teilnahme bei der jährlich stattfindenden internationale Konferenz *Winter Conference on Applications of Computer Vision* (WACV) möglich, welche sich mit dem aktuellen Stand der Forschung zu verschiedenen Anwendungsgebieten der computergestützten Bildverarbeitung und -analyse sowie zugehöriger Themen wie Datenmanagement und -aufbereitung, Objekt- und Personenerkennung, *Deep Learning* und Clustering beschäftigt und vom 7.-11. Januar 2019 in Hawaii, USA stattfand. Die auf den während des TRECVid-Wettbewerbs erzielten Erkenntnisse konnten hier erweitert einem breiteren und internationaleren Forschungs- und Fachpublikum präsentiert, sowie die Sichtbarkeit des Forschungsprojektes localizeIT vergrößert werden. Der Beitrag zu TRECVid-Wettbewerb wurde hier vorgestellt, wobei der besondere Schwerpunkt hierbei auf der Vorstellung der auf Virtualisierungs- und Webtechnologien basierenden Infrastruktur zur Kontrolle, Steuerung und Skalierung der angeschlossenen Detektions-, Analyse-, Auswertungs- und Darstellungsinfrastruktur für die Suche von Personen, Objekten und Ereignissen in großen Datenmengen lag, entsprechend Abbildung 16.

Viele der bei der Konferenz vorgestellten wissenschaftlichen Arbeiten befassten sich mit dem Thema der Verarbeitung, Analyse und Interpretation von visuellen Daten und darauf aufbauender Aktionsauslösung, was die große Bedeutung dieses Themas für die internationale Wissenschaft, aber auch für den wirtschaftlichen Einsatz verdeutlicht. Von speziellen Interesse für das Projekt ist dabei *LUCFER: A Large-Scale Context-Sensitive Image Dataset for Deep Learning of Visual Emotions* von der University of Central Florida, welches das Erstellen einer Datenbank für die Erkennung von Emotionen, deren Erlernung durch automatisierte Systeme ermöglicht bzw. erleichtert. Zum Anderen der Beitrag der University of Buffelo bezüglich den Space-Time Event Clouds for Gesture Recognition: From RGB Cameras to Event Cameras. Dies stellt interessante Herangehensweisen für den Bereich der Verhaltenserkennung des Projektes dar. In Kombination mit der Analyse des optischen Flusses wie beispielsweise in *A Fusion Approach for Multi-Frame Optical Flow Estimation* der Georgia Tech. Die durch die Teilnahme an der WACV-Konferenz erzielten Forschungsergebnisse stellen gute Lösungen für den Einsatz im Projekt und zur Optimierung bereits vorhandener projektrelevante Lösungen dar. So können die Bewegungen im Aufnahmebereich der Kameras mittels des optischen Flusses verfolgt und als komplexere Abläufe in der Verhaltens- und Gestenerkennung genutzt werden. Darüber hinaus konnte die Sichtbarkeit des Projektes localizeIT, seiner Stifter und der InnoProfile-Transfer-Initiative verbessert werden.

[5] `http://www.viratdata.org`

Das Abschneiden im TRECVID-INS-Wettbewerb endete in der Regel auf den hinteren Plätzen, wobei die erzielten Ergebnisse Personen an bestimmten Orten zu finden zwischen knapp 10% und fast 60% *average precision* schwankten. Vor dem Hintergrund des immens großen Datensatzes sind diese Ergebnisse für das Projekt zufriedenstellend. Die personellen und rechentechnischen Ressources des Projekts reichten erwartungsgemäß nicht aus um an Top-Playern vom MIT (Massechusetts Institute of Technology) oder IBM in den Rankings vorbeizuziehen. Angesichts der Komplexität des INS-Wettbewerbs ist die Teilnahmehürde so hoch, dass viele Gruppen gar nicht teilnehmen. So ist es auch als Erfolg zu verbuchen, dass im Rahmen von LocalizeIT ein deutscher Vertreter regelmäßig im Wettbewerb vertreten war. Zu den technischen Details und Algorithmenbeschreibungen sei der geneigte Leser auf die jeweiligen *Working Notes* zum Wettbewerb verwiesen [67, 66, 121, 124, 123].

2.2.8 Objekt/Personenerkennung/tracking III: Analyse und Evaluation von Algorithmen in Laborszenarien

Das in Kapitel 2.2.2 vorgestellte und in Abbildung 17 dargestellte Labor wurde in diesem Arbeitsbereich zur Erstellung systematischer und reproduzierbaren Mehrkamera-Video-Aufnahmen genutzt. Unter kontrollierten Bedingungen und Beleuchtungssituationen entstanden verschiedene Szenarien mit visuellen Sensoren wie dem smarten S2000 Stereo-Sensor, teils mit gleichzeitigem Einsatz der Mikrofon-Arrays. Details zu den Szenarien sind der Datensatzübersicht in Kapitel 2.2.3 zu entnehmen. Gleichzeitig wurden einige Sensoren für die Aufnahme verschiedener Szenen außerhalb von Gebäuden eingesetzt und somit Datensätze mit größerer Varianz und mehr Störeinflüssen wie Hintergrundrauschen, unterschiedlichste Licht- und Wettersituationen etc. aufgezeichnet, Abbildung 18. Auf diese Weise wurden grundlegende Erkenntnisse des TRECVID-Wettbewerbs, dargestellt in Kapitel 2.2.7, erweitert zur qualitativen und quantitativen Beurteilung der entwickelten und angewandten Systeme im Umfeld realer, aber auch für die Stifter relevanter Szenarien.

Durch die genaue Auswahl der Szenarien mit dominanten, gezielt gewählten Objekten und z.T. unbekannten Randparametern ist zum Einen eine Untersuchung der Leistungsparameter, Funktionsgrenzen und Eigenschaften der entwickelten und angewandten Systeme möglich, zum Anderen aber auch eine Nutzung für das mit dem Kooperationspartner Intenta abgesprochene Einsatzgebiet der Linienbusinnenaufnahmen.

Personenzählung

Zur automatisierten Erfassung der aktuellen Fahrgastsituation in öffentlichen Verkehrsmitteln mit vorhandenen 2D Videokameras wurde untersucht, inwiefern sich das Detektionsframework OpenPose zur Personenzählung eignet. Die Aufnahmen wurden mit einer

bzw. 3 Kameras durchgeführt, siehe Abbildung 19. Die Szene beinhaltet einen Sitz für den Busfahrer sowie 11 Sitzplätze für die Fahrgäste. Von den einzelnen Kameras wird in diesem Szenario meist nicht das gesamte Verkehrsmittel erfasst und es können Verdeckungen einzelner Personen auftreten. Die zur Analyse erhobenen Rohdaten bestehen aus 2 Videoaufnahmen, in denen typische Fahrgastsituationen nachgestellt wurden. Die erste Aufnahme erfolgte mit Kamera 3 und beinhaltet 3728 Videoframes. Die zweite Aufnahme wurde mit allen 3 Kameras gleichzeitig ausgeführt und beinhaltet bis zu 1780 Videoframes pro Kamera. Die Videoaufnahmen wurden per Hand annotiert und auf die eigentlichen Busszenen beschränkt. Für jede Kameraperspektive erfolgte die Festlegung der Fläche des Businneren, siehe Abbildung 20 und für jeden Videoframe wurde die Anzahl der Personen im Bus und die Anzahl der Personen innerhalb der Fläche des Businneren festgehalten, siehe Tabelle 4. Um eine später angedachte Sensorfusion zu ermöglichen, erfolgte eine Synchronisierung der Daten von Aufnahme 2 auf eine einheitliche Videoframeanzahl von 1685 für alle 3 Kameraperspektiven. Zur Untersuchung, wie gut OpenPose die genaue Anzahl der Personen im Bild bestimmen kann, wurde das Fehlermaß $RootSquaredError$ genutzt. Dies ist in unserem Fall die Wurzel der quadrierten Differenz der exakten Personenzahl und der detektierten Personenzahl in einem Bild. Um einen Durchschnittswert für die gesamte Videosequenz zu erhalten, werden die Fehler über alle Videoframes gemittelt. Berücksichtigt werden hierbei nur die Personen innerhalb des Bussbereiches (AOI) der jeweiligen Kameraperspektive, siehe Abbildung 20. Je nach gewähltem $Score_{Min}$-Wert, ab dem eine Person als detektiert gilt, ergibt sich ein anderer $RMSE_{AOI}$-Wert. Für die Ermittlung des minimalem $RMSE_{AOI}$-Wertes wurde ein Greedy-Algorithmus, siehe [105], S. 161, entwickelt, der den jeweils optimalen $Score_{Min}$-Wert ermittelt. Die Ergebnisse des Greedy-Algorithmus sind in den Abbildungen 21 und 22 visualisiert und die jeweils optimalen $Score_{Min}$-Werte in Tabelle 5 aufgelistet. Aus diesen Ergebnissen folgt, dass OpenPose zur Erfassung der Fahrgastsituation geeignet ist. Je nach Kameraperspektive und Personenverdeckung schwankt der durchschnittliche Fehler bei der Personenzählung von 0.15 bis 0.84 Personen pro Bild.

Aufnahme 1	Aufnahme 2		
max. 6 Personen im Bus	max. 4 Personen im Bus		
Kamera 3	Kamera 1	Kamera 2	Kamera 3
3557 Frames	1685 Frames	1647 Frames	1550 Frames

Tabelle 4: Resultate der Videoaufbereitung und Annotation

Mittels dieses auf der *International BCS Human Computer Interaction Conference* (BHCI) veröffentlichten Lösungsansatzes [87] sind etliche Parameter wie Perspektiven und oder Helligkeitsänderungen einfacher zu testen, ohne den Aufwand größerer Hardwareanpassungen. Eine Verifizierung der durch synthetischen Daten gefundenen Lösungen erfolgt an-

	Aufnahme 1	Aufnahme 2		
	Kamera 3	Kamera 1	Kamera 2	Kamera 3
$Score_{Min}$	0.0753	0.0199	0.0442	0.0660
$MRMSE_{AOI}$	0.5545	0.8386	0.1481	0.3258

Tabelle 5: Minimal Root-Mean-Squared Errors ($MRMSE_{AOI}$)

schließend im Labor. In Kooperation mit Forschern an der Hochschule Mittweida konnte weiterhin ein Studentenprojekt gewonnen werden, welches eine Weiterentwicklung der Lösung für Bewegungsdetektionen mit Smartsensoren untersucht.

Fahrraderkennung aus der Vogelperspektive
Die Erkennung von Fahrrädern von oben stellte für die Objektererkennungs-Frameworks Detectron ([44]) und YOLO (You Only Look Once [96]) eine große Herausforderung dar. Diese Frameworks sind auf Objekterkennung trainiert mit Bilddaten, die im Wesentlichen frontal aufgenommenen worden sind. Daher musste eine Anpassung vorgenommen werden. Die Einbindung der in Kapitel 2.2.3 vorgestellten Fahrrad-Aufnahmen in die Trainingsdaten mit Fahrrädern in unterschiedlichen Ausführungen und Perspektiven ergab eine deutliche Verbesserung der Detektionseffizienz in der Vogelperspektive. Die über Transferlernen neutrainierten Modelle von YOLO in Version 2 und 3, sowie SSD (singe shot detector, [83]) und Faster RCNN ([98]) ergaben eine *mean Average Precision* mAP von 81% für das FRCNN-Modell für die Klassifikation der Laborvideoaufnahmen.

2.2.9 Objekt/Personenerkennung/tracking IV: Animation virtueller Szenarien zur Evaluation von Algorithmen

Eine synthetische Nachbildung eines Linienbusses mit Fahrer und Passagier wurde erstellt, welche auf dem im Rahmen der wissenschaftlichen Beiträge [87] und [88] vorgestellten bzw. weiterentwickelten Frameworks *SyntTEV* beruht.

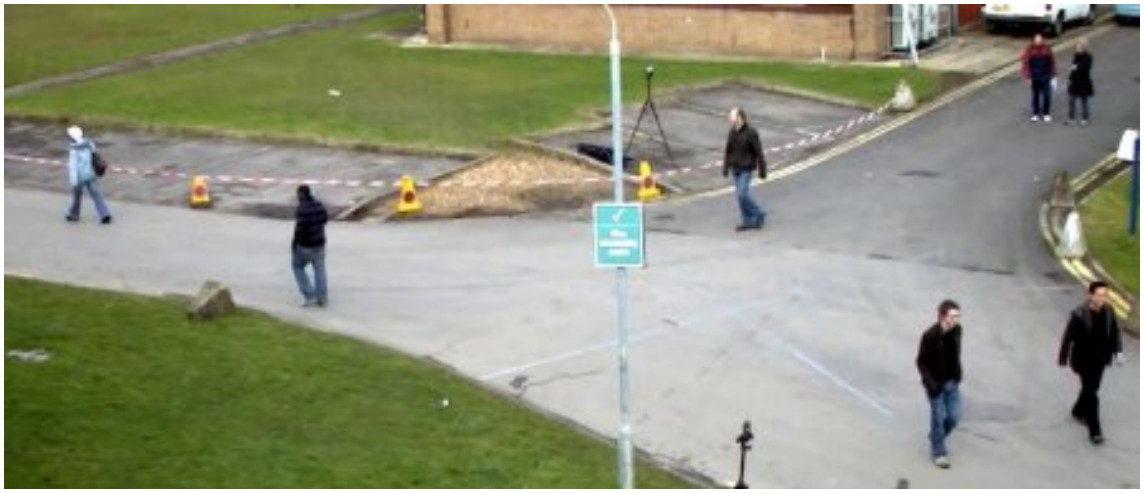
Vor diesem Hintergrund wurde weiterhin eine synthetische Nachbildung eines Linienbusses mit Fahrer und Passagier erstellt, Abbildung 26, welche auf dem im Rahmen der wissenschaftlichen Beiträge [87] und [88] vorgestellten bzw. weiterentwickelten Frameworks *SyntTEV* beruht. Mit diesem konnten verschiedene Aspekte und Eigenschaften der dafür verwendeten Systeme im Rahmen einer Software-in-the-Loop Begutachtung frühzeitig und während der Entwicklung erprobt und verbessert werden. Sind zukünftige Anpassungen am Szenario vorzunehmen, so können deren Auswirkungen schnell und kosteneffizient mit der Nachbildung untersucht und erste Optimierungen durchgeführt werden bevor aufwendigere und teurere Realweltuntersuchungen für die Endoptimierung erfolgen.

Weiterhin waren Einreichungen bei der internationalen Konferenz *International Conference on Intelligent Human Systems Integration* (IHSI) erfolgreich, welche vom 7.-10. Februar 2019 in San Diego, USA, stattfand. Hier stellt der Bereich der Detektionen und Analysen von Personen, Objekten und Ereignissen in großen Videodaten einen besonders interessanten Teilbereich dar. Dazu wurde ein System vorgestellt um die Leistungsfähigkeit moderner Bild- und Videoanalysesysteme zur Posen-, Personen- und Objektdetektion von möglichst vielen Betrachtungsrichtungen, unter möglichst vielen Beleuchtungsvarianten von möglichst vielen unterschiedlichen Personen mit unterschiedlichsten Bekleidungen zu beurteilen. Wie in Abbildung 27 dargestellt, erfolgt die hierbei die Festlegung der Eigenschaften von Objekten, Personenmodellen und Aktivitäten sowie deren entsprechenden Umgebung. Diese werden mit der Open-Source 3D-Modellierungs- und Rendersoftware Blender1 zu Szenarien mit den für die Detektion gewünschten Eigenschaften kombiniert. Alle Eigenschaften sind somit bekannt, fehlerfrei und generierte Bilddaten benötigen keine Annotation. Darüber hinaus können so nahezu unbegrenzte Mengen an Trainings- und Testdaten erstellt werden. Im Zusammenspiel mit dem im Projekt localizeIT aufgebauten Labor und der maßstabsgetreuen Nachbildung als Szenario ist eine umfangreiche und realitätsnahe Überprüfung der Detektionsleistungen, speziell bei Bewegungsabläufen, möglich.

Weiterhin war hier ein Beitrag zum Promothionsthema einreichbar, welches sich mit der Aufnahme, Verarbeitung und Wiedergabe von Bild- und Videodaten bei technischen Systemen mit rechteckige bzw. quadratische Formen für die Bildpunkte (Pixel) und den biologischen, über Jahrmillionen optimierten Systemen wie Facettenaugen oder der Netzhaut von Linsenaugen und ihrer fast ausschließlich sechseckige Formen beschäftigt. Da verschiedene Untersuchungen zu diesen Formunterschieden deutliche Vorteile für die hexagonale Form aufweisen aber auf das Fehlen nativer Aufnahmesysteme als größtes Hindernis für genauere Untersuchungen hervorhob, wurde ein System zur Lösung vorgestellt. [85] Dies ermöglicht weitere Forschungen und Untersuchungen ohne die kostenintensive und langwierige Entwicklung von hardwarebasierten Aufnahmesystemen. Die entwickelte virtuelle Kamera und erste Ergebnisse wurden im Rahmen der Konferenz dem Fachpublikum vorgestellt, wie in Abbildung 28.

Der vorgesehene 2. Workshop wurde mit im Rahmen eines eigenen Workshops bei der Veranstaltung *Chemnitzer Linux Tage 2019*[6] durchgeführt und die einzelnen Ergebnisse dem wissenschaftlichen Fachpublikum, sowie den meist technisch interessierten Veranstaltungsteilnehmern präsentiert.

[6] `https://chemnitzer.linux-tage.de/2019/de/programm/workshops`



(a) Beispielszene für Vordergrund-Hintergrund-Trennung mit mehreren, sich bewegenden Personen als Vordergrund.

(b) Ergebnis einer Vordergrund-Hintergrund-Trennung mit weißen Markierungen für Vordergrundobjekte, graue Markierungen für Bereiche unterschiedlicher Konfidenz der Zuordnung zum Vordergrund und schwarzen Hintergrundbereichen. Die Klassifizierung des rot-weißen Absperrbandes als Vordergrund stellt dabei einen typischen Fehler aufgrund seiner zu starken Bewegung in der Szene dar.

Abbildung 12: Beispielszene mit dem Ergebnis einer Vordergrund-Hintergrund-Trennung.

Abbildung 13: Das in dieser Videoszene unmittelbar vorher verlorene Papierbündel (roter Kreis) ist für Menschen relativ gut erkennbar. Automatisierte Trennungsverfahren können aufgrund der Ähnlichkeit zum Untergrund allerdings nur sehr schlecht zwischen beidem unterscheiden.

Abbildung 14: Das linke Bild zeigt ein Beispiel mit der Maske des zu findenden Objektes, der Frau namens Dot und im rechten Bild einen Ort. Korrekte Erkennungen im Datensatz sind alle Bilder in denen Dot an diesem Ort zu erkennen ist.

Abbildung 15: Beispielaufnahme aus dem VIRAT-Datensatz zeigt einen Parkplatzes mit verschiedenen Objekten und Personen, deren Annotation ihre Tätigkeit beschreibt.

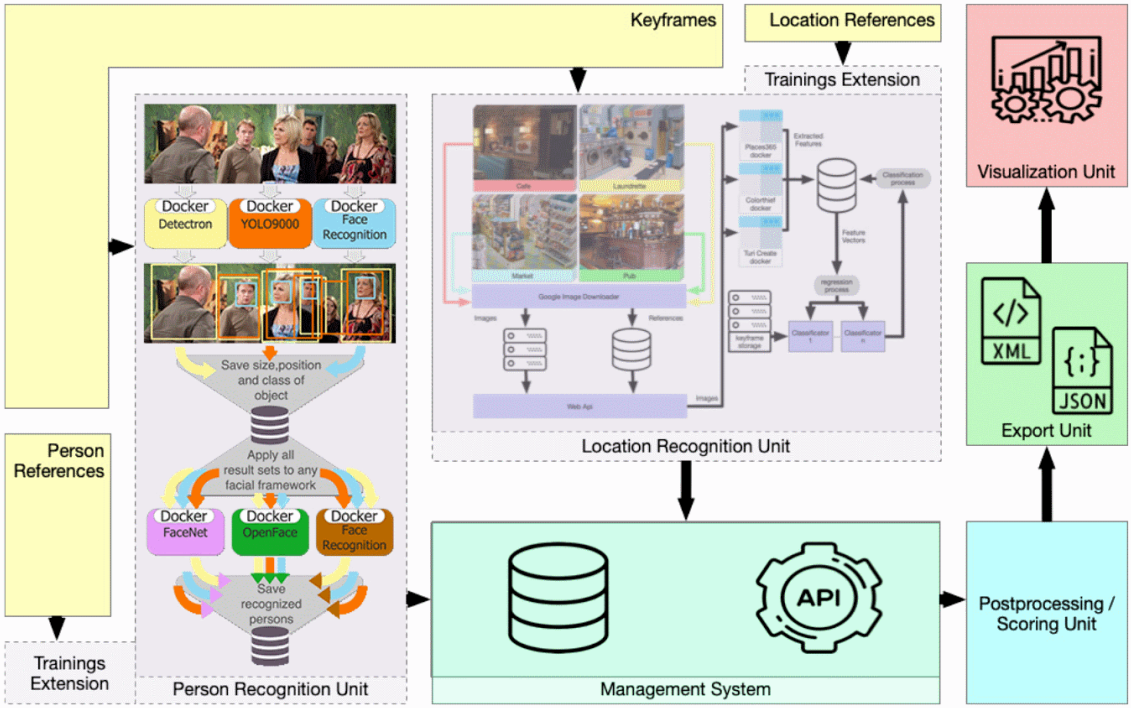
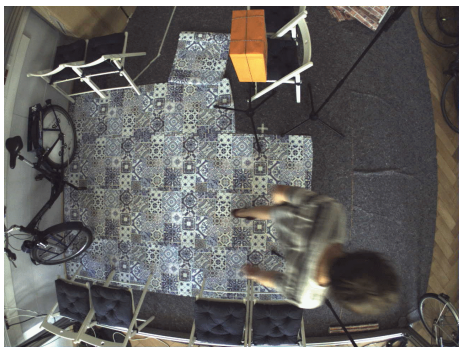


Abbildung 16: Schematischer Aufbau der virtualisierten Datenanalyseinfrastruktur mit Docker-basierten Detektionssystemen, Datenbank und Ergebniserstellungs- und Darstellungsoberfläche. [122]

(a)

(b)



(c)

Abbildung 17: Exemplarische Beispiele der Audio-Video-Labornutzung zur Datensatzerstellung: (a) Beispielaufnahme des Labors mit regelmäßig verteilten Lautsprechern.
(b) Innenaufnahme eines nachgebildeten Linienbuses mit Eingangsbereich, Fahrer, Ticketentwertungsautomat und mehreren Sitzplätzen.
(c) Innenaufnahme eines nachgebildeten Linienbuses im Bereich der Fahrradstellplätze mit Eingangsbereich, Ticketentwertungsautomat und mehreren Sitzplätzen, einem Fahrrad und modifiziertem Untergrund.

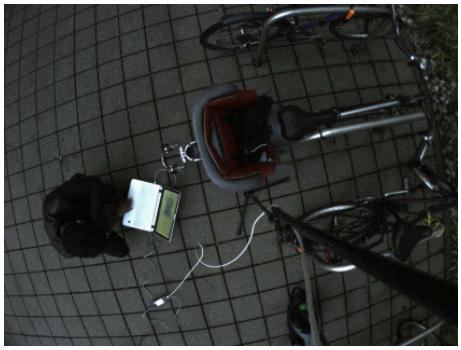


Abbildung 18: Beispielaufnahme der Datensammlung der außerhalb des Labors aufgenommenen Daten

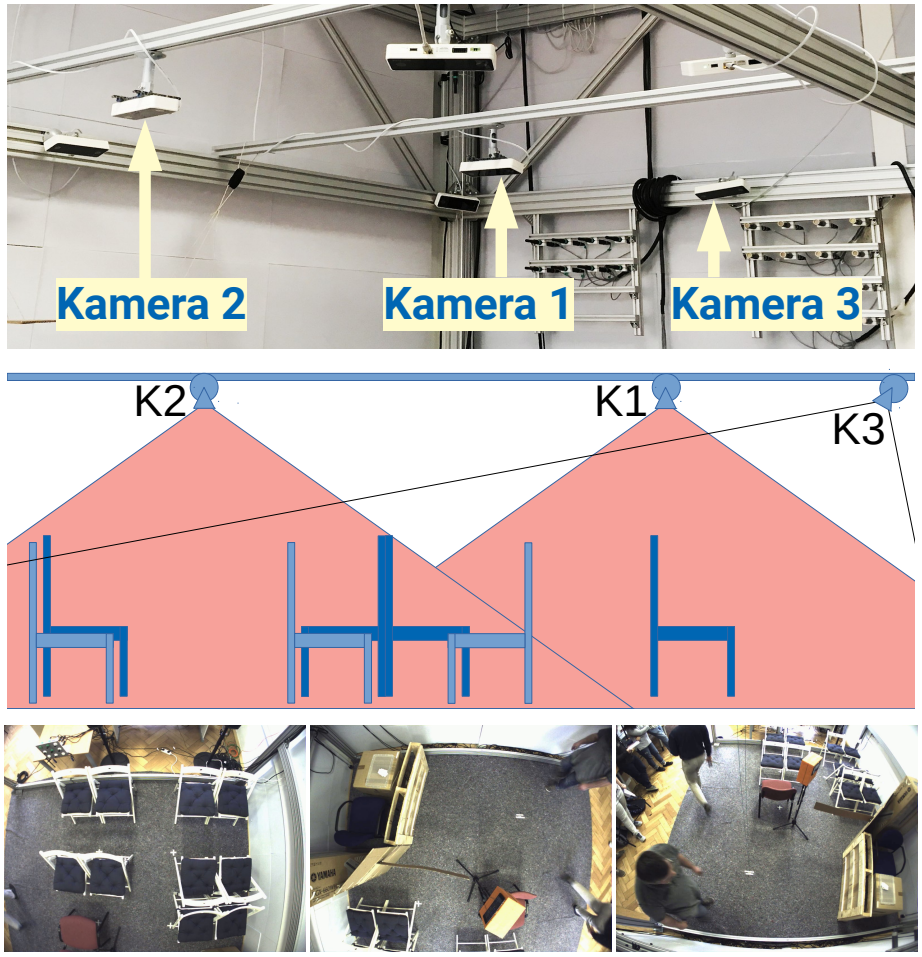


Abbildung 19: Laboraufbau für die Busszene mit 3 Kameras

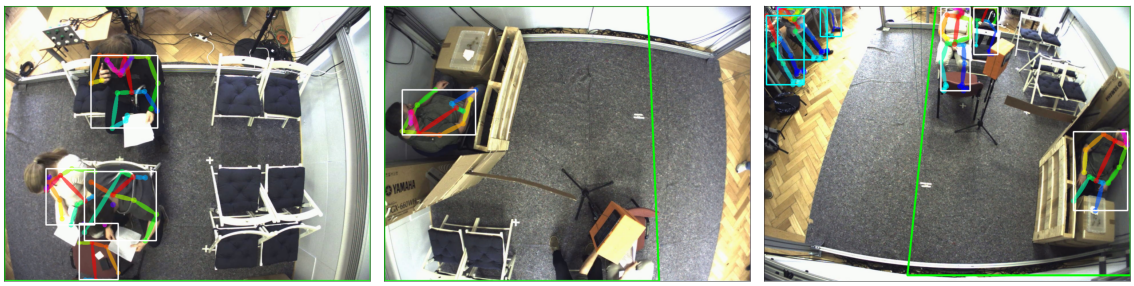


Abbildung 20: Festlegung des Businneren für jede Kameraperspektive, hier dargestellt mit einer grünen Umrandung, für die Kameras 2 (links), 1 (Mitte) und 3 (rechts). Alle mit OpenPose detektierten Personen sind mit einer Boundingbox umrandet (weiß für im Bus befindlich, türkis für außerhalb) und die jeweils detektierte Skelettstruktur wurde mit eingezeichnet.

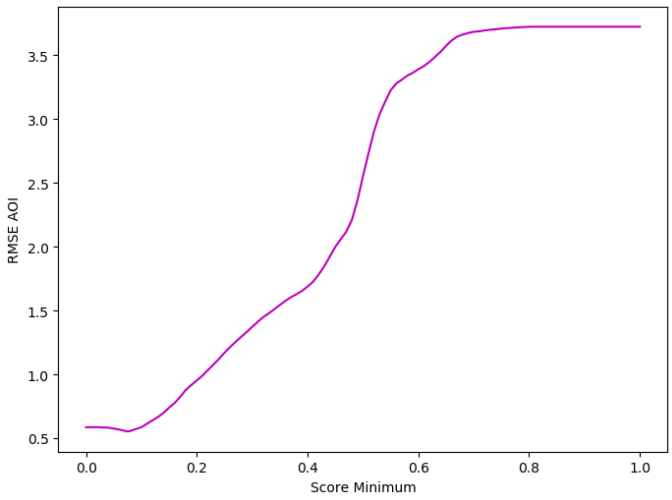


Abbildung 21: Root-Mean-Squared Errors ($MRMSE_{AOI}$) für Aufnahme 1.

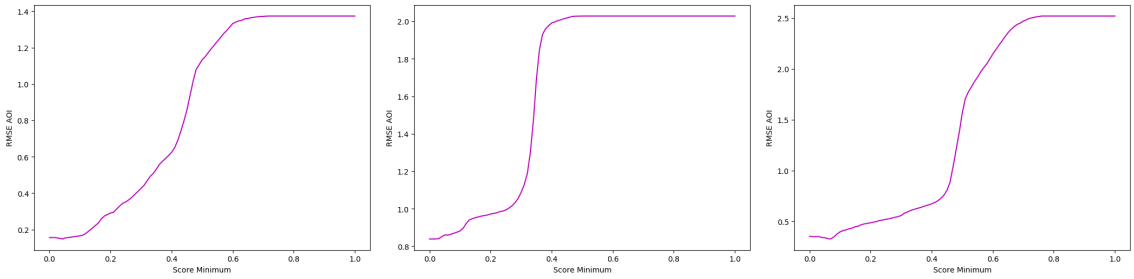


Abbildung 22: Root-Mean-Squared Errors ($MRMSE_{AOI}$) für Aufnahme 2, für die Kameras 2 (links), 1 (Mitte) und 3 (rechts).

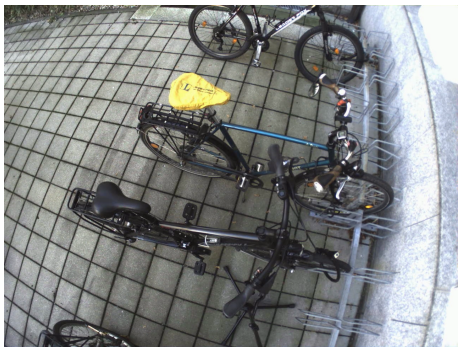


Abbildung 23: Beispielaufnahme der Datensammlung zur Erkennung von Fahrrädern

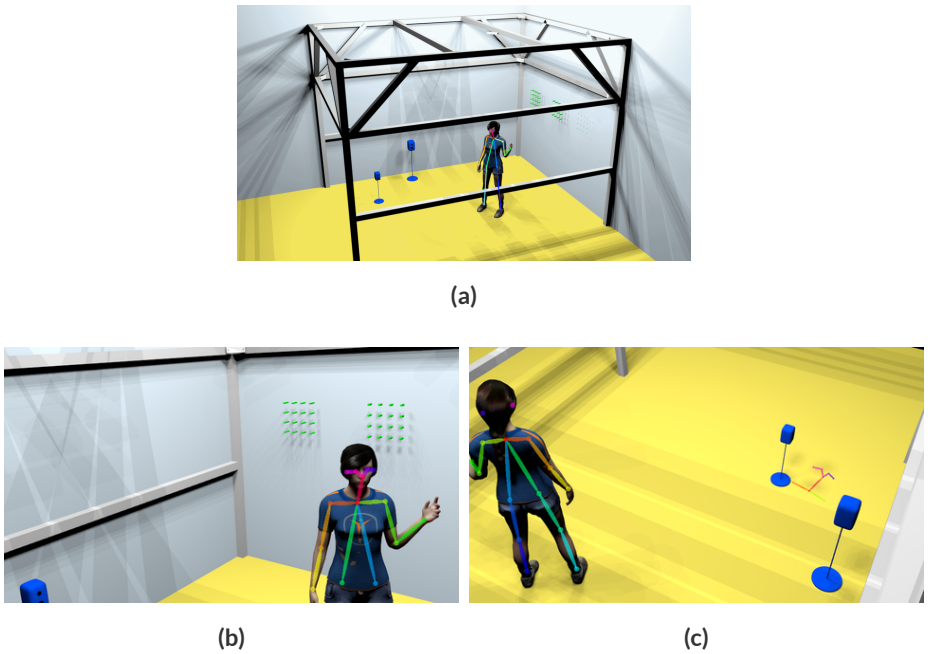


Abbildung 24: Beispiele des virtuellen Labors mit Aufnahmen von verschiedenen Blickwinkeln und den zugehörigen Detektionen inklusive Fehldetektionen durch Schattenwurf. [87]

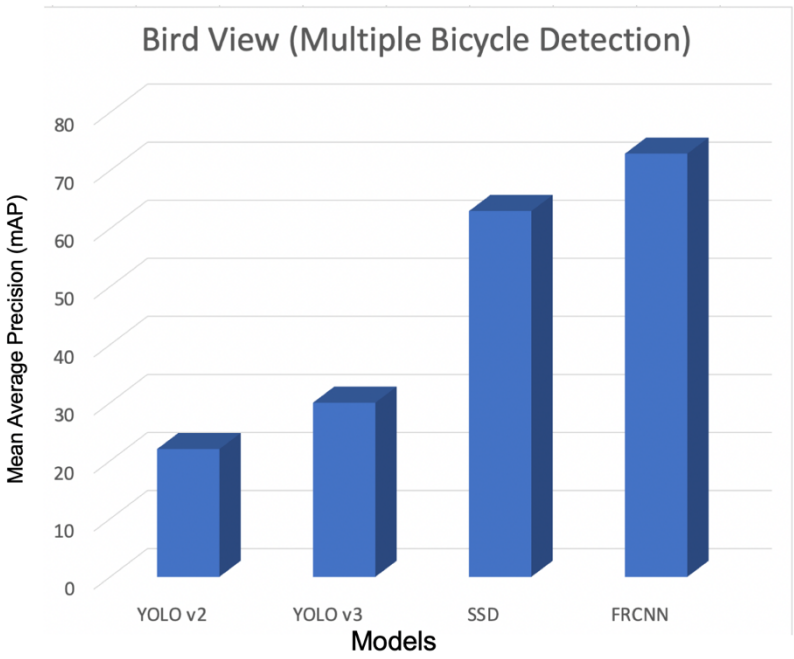


Abbildung 25: Ergebnisse der Fahrraderkennung mit unterschiedlichen Modellen.

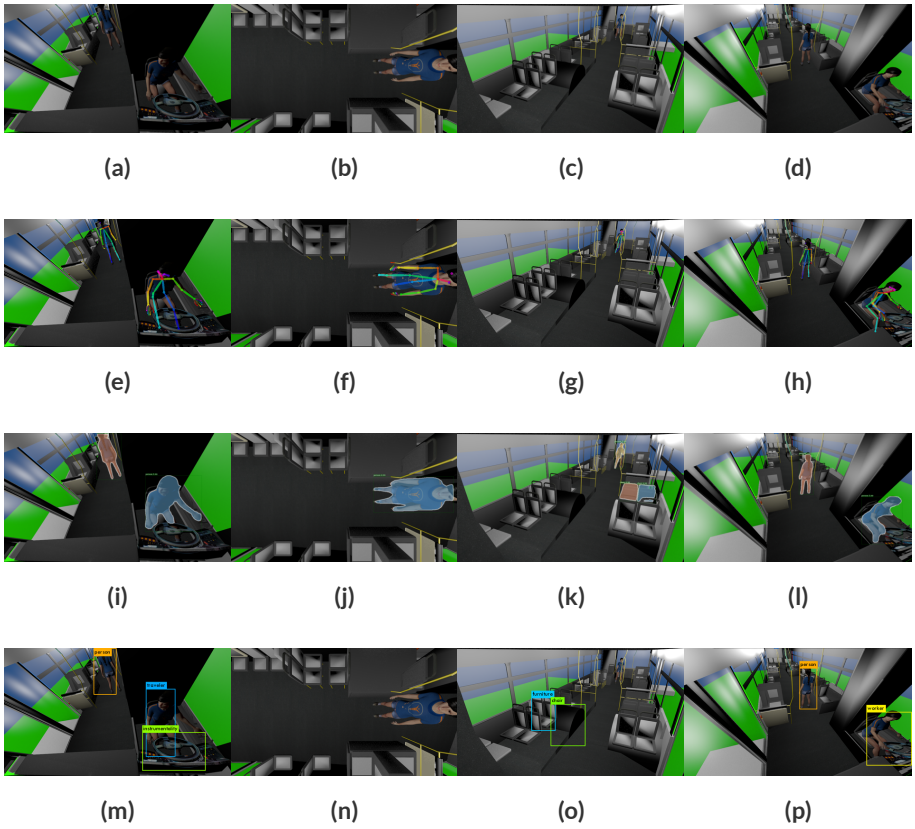


Abbildung 26: (a-d) Aufnahmen von ausgewählten Bereichen des virtuellen Busses. (e-h) Detektionen von Openpose zur Lokalisierung von Personen und ihren Körperteilen. (i-l) Untersuchung der Aufnahmen mittels Detectron zur Erkennung der Umgebung. (m-p) Zur Verifikation folgt die Anwendung von Yolo9000.

Die Erkennung von Personen zeigt sich nahezu Fehlerfrei und Präzise, wohingegen beispielweise Sitze trotz guter Sicht nahezu komplett unbeachtet bleiben.

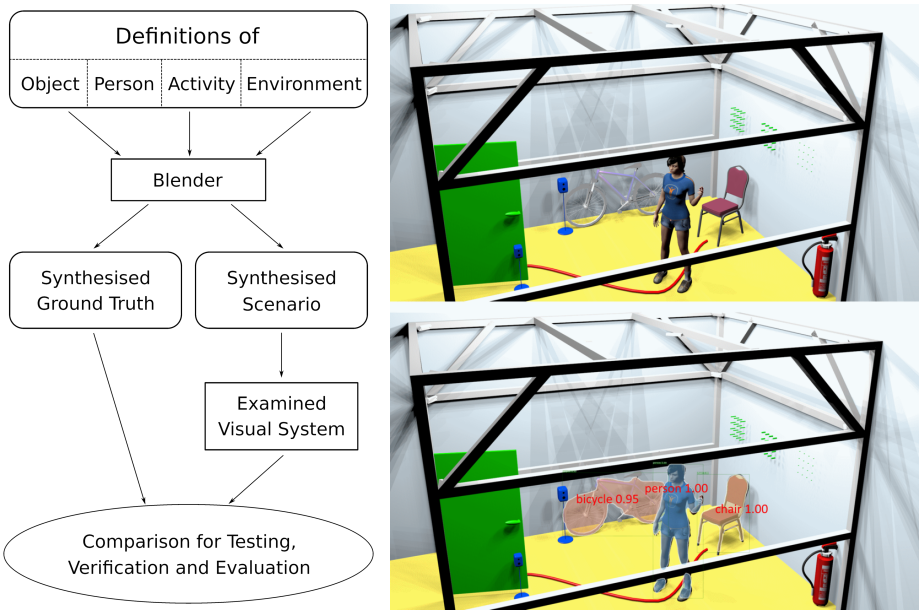


Abbildung 27: Links)Schematischer Ablauf zur Generierung von Daten für die Analyse visueller Detektions- und Analysesysteme. Rechts oben) Beispiel des virtueller Nachbaus des localizeIT-Labors mit Gegenständen, Person und Handgeste. Rechts unten) Ergebnisse mit Detektionstyp und zugehörigem Konfidenzwert. [88]

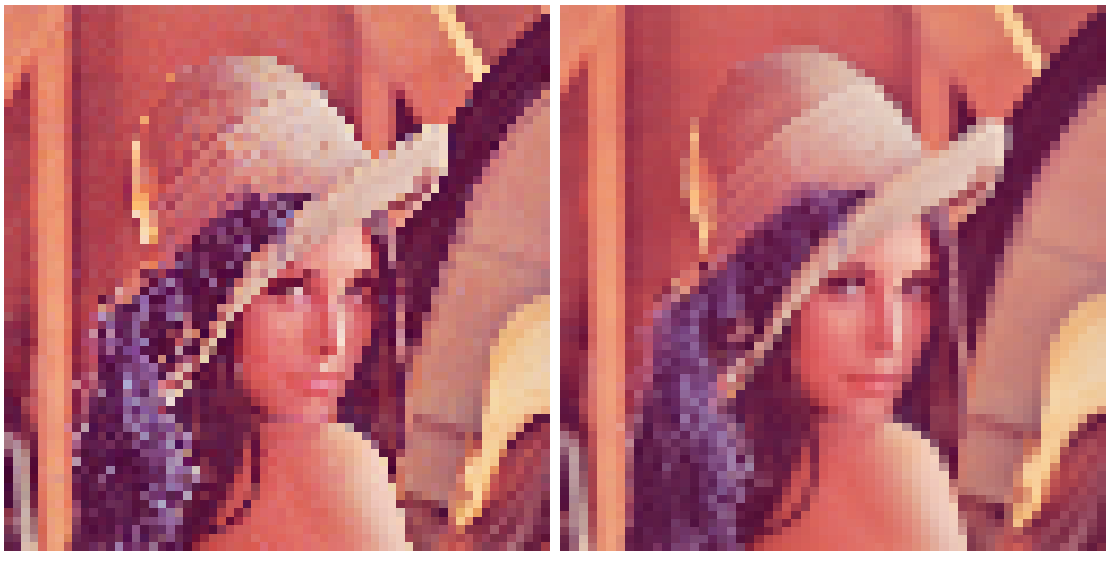
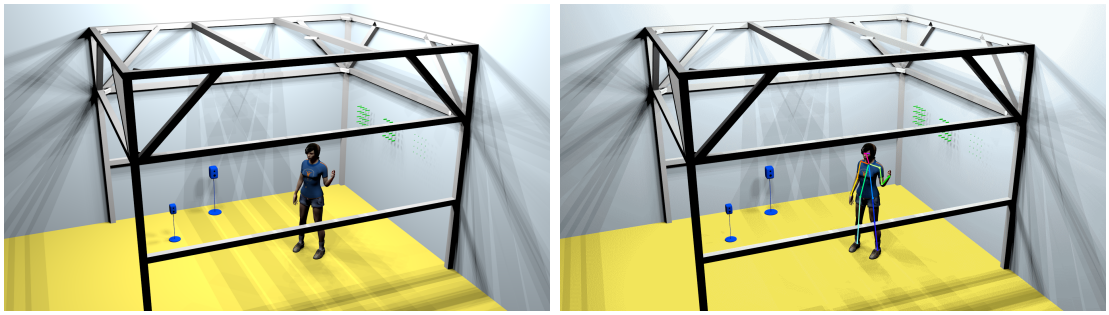


Abbildung 28: Testbild Lena mit (a) quadratische und (b) hexagonalen Pixeln gleicher Fläche. [85]



(a) Labor (b) Labor Openpose

Abbildung 29: Synthetische Umgebung mit Darstellung des Labors und der Detektionenergebnisse von Openpose.

2.2.10 Bilanz nach Erreichen des letzten Meilensteins

Komplementär zu den im Labor erstellten Szenarien, die das Verhalten von Menschen in öffentlichen Verkehrsmitteln ohne und mit Fahrrädern nachstellen, wurde vor allem der Ansatz virtueller Umgebungen eruiert. Dazu hat sich das Programm Blender als probate Simulationsumgebung erwiesen um Varianz skriptbasiert z.B. in Python zu erzeugen, die sich in Laborumgebungen nur mit erheblichem Aufwand schaustellerisch nachbilden lässt. Systematisch ändern lassen sich u.a. Kamera-Sensoren und Personen in Häufigkeit und Position, aber auch Einflüsse durch Bewegung, Licht und Verdeckung durch Objekte. Vorteilhaft ist, dass hier auf manuelles Annotieren weitgehend verzichtet werden kann, da die räumliche Lage von Personen und Objekten sowohl im 3D Raum als auch in der 2D-Projektion im Kamerabild bekannt ist. Damit ergänzen Blender-Animationen die vorhandenen Labor-Videos. Auf dem Laborvideomaterial konnte durch die Kombination mehrerer auf CNN-basierter Algorithmen und mehrerer Sensoren zur Personenerkennung gezeigt werden dass eine intelligente Fusion durch Entscheidungsbaumoptimierung eine deutliche Verbesserung der Personenzählung ganz ohne Tracking ermöglicht. Das Einbeziehen von Tracking zur Verhaltensanalyse konnte im Rahmen der ActEv Challenge im Rahmen des TRECVID-Wettbewerbs 2019 prototypisch gezeigt werden [124]. Bis zum Projektende sollen die Verfahren unter Zuhilfenahme von Blender verfeinert und abschließend evaluiert werden.

2.3 AB 2: Lokalisierung und semantische Verknüpfung von Bilddaten

Bildagenturen haben einen täglichen Bilddurchsatz im oberen 5-stelligen Bereich. Bei bestimmten Ereignissen kommen sie schnell auf über 100.000 Bilder pro Tag. Eine solche Menge lässt sich nicht mehr sinnvoll händisch mit Beschreibungen versehen. Hier sind automatische Verfahren notwendig, die eine konkrete Lokalisierung der Aufnahme möglich machen. Typisch ist beispielsweise die Aufnahmesituation roter Teppich bei der Berlinale. Hier werden in kürzester Zeit zahlreiche Photographien gemacht. Mit modernen Kameras kommen die Bilder automatisch mit GPS-Koordinaten und Aufnahmezeitpunkt in die Datenbank, aber ohne weitere semantische Beschreibung. Wenn nun Verlage im Bildarchiv recherchieren, werden Sie weder die GPS-Koordinaten, noch den genauen Aufnahmezeitpunkt des gewünschten Bildes kennen. Hier können zwei Methoden helfen: 1) Webservices stellen eine Verbindung zu Ereignisdatenbanken sowie Geoinformationen her. So kann zumindest die Aufnahmesituation Berlinale ermittelt werden. 2) Für die weitere semantische Analyse des Materials müssen Verfahren der Bilderkennung herangezogen werden, um beispielsweise Pose oder Blickrichtung einer auf dem roten Teppich fotografierten Person zu erkennen. Im folgenden Unterkapitel werden zunächst die wichtigsten Aufgaben und Ziele aus dem Originalantrag zusammengefasst und deren Umsetzung im Projekt kurz skizziert. Im darauffolgenden Unterkapitel werden die jeweiligen Ergebnisse detailliert besprochen.

2.3.1 Aufgabenstellung und Zielsetzung

Im Projekt wurden Aufgaben quartalsweise formuliert und auf diese Weise auch in den Zwischenberichten abgehandelt. In der folgenden Auflistung sollen jedoch eher die übergeordneten Inhalte betrachtet und deren Lösung kurz angerissen werden:

1. Analyse von Perspektive, Blickwinkel und Pose:
   Die Perspektive der Kamera, sowie die Pose und der Blickwinkel von Personen relativ zur Kamera sollten automatisch bestimmt werden, um als Metadaten und damit als Auswahlkriterien zur Verfügung zu stehen. Genutzt wurden dabei die in Kapitel 2.2.3 eingehender vorgestellten Datensätze mit Rotem-Teppich-Szenario, im Labor nachgestellte Busszenarien und die in Blender erstellte Szene einer laufenden Person mit Fahrrad. Im Projektkontext wurde die Perspektive als etwas betrachtet, das häufig gleich und gegeben ist. Daher ging es nicht um die algorithmengestützte Bestimmung der Kameraperspektive relativ zu Objekten, sondern vermehrt darum, die Bestimmung aller sonstigen Metadaten unter kontrollierten Kameraperspektiven zu untersuchen oder gar die Detektion anzupassen. So entstanden bspw. eigene Klassifikatoren zur Fahrraderkennung aus der Vogelperspektive. In Realvideos aber auch

in virtuellen Szenarien wurde die Winkelabhängigkeit der Kamera relativ zum Objekt untersucht, vor allem in der Gelenkpunkterkennung (*human pose estimation*), der Objekterkennung (Fahrrad), der Kopferkennung und der Personenzählung. Für die Bestimmung des Blickwinkels wurden bestehende Algorithmen der Kopfposenerkennung gegeneinander evaluiert. Ziel war es, Informationen zu gewinnen, ob Personen in Richtung Kamera schauen. Menschliche Posen standen im Mittelpunkt der Untersuchungen im Rahmen der Verhaltenserkennung im TRECVID-Wettbewerb (siehe Kapitel 2.2.7) und in der Analyse des Datensets mit Personen auf einem roten Teppich. Leistungsstarke Algorithmen wurden evaluiert, die die menschliche Pose als Skelett mit sogenannten *Keypoints* annähern. Dazu gehören neben Gelenkpunkten wie Knien, Ellbogen, Schultern, Nacken, etc. auch Gesichtsmerkmale wie Augen, Nase und Mund. Die Gesichtspunkte konnten auch genutzt werden, um etwas über die Kopfpose –also den Blickwinkel –auszusagen.

2. Verortung von Bilddaten mit GPS mit Ortsnamen:
   Der Fachbegriff zu diesem Thema ist *Reverse Geocoding* und behandelt das Abrufen von Ortsinformationen aus GSP-Daten, die heute oft in den Bild-Metadaten gespeichert sind, wenn bei der Aufnahme die entsprechende Funktion aktiviert wird. Untersucht und evaluiert wurden 3 verbreitete Webservices auf Feldern wie Land, Stadt, Straße und sog. *Points of Intersts*. Letztere stellen relevante Informationen für die automatische Verschlagwortung dar, z.B. ob sich in der Nähe der Koordinate relevante Lokalitäten befinden wie Restaurants, Tankstellen, Sehenswürdigkeiten, etc.

3. Verortung GPS-loser Bilddaten mit Ortsnamen:
   Ein weiterer im Projektkontext relavanterer Fall sind Bilder ohne GPS- und Ortsinformationen, die ebenfalls mit Ortsinformationen zu verschlagworten sind. Diese Aufgabe wurde eingehender im TRECVID-Wettbewerb untersucht, im Rahmen der *Instance Search Tasks*, bei denen Personen an bestimmten Orten gefunden werden müssen. Dazu wurden zum Einen CNNs auf Orte wie Bars, Wohnzimmer oder Wäschereien trainiert aber auch bestehende Algorithmen herangezogen, die Orte in Fotos klassifizieren können, wie z.B. Vorlesungssaal, Wald, Strand, Cafeteria, Schlafzimmer usw.

4. Abgleich von Orten und Ereignissen/Metadatenanreichung mit Ereignisdaten:
   Unter der Nutzung des Zeitstempels fotografischer Aufnahmen war die Aufgabe Ereignisdatenbanken zu nutzen, um Bilder auch mit Ereignis-Metadaten zu verschlagworten, z.B. Berlinale 2017 oder Immatrikulationsfeier der TU Chemnitz 2019. Dazu wurde eruiert, welche Formen oder Quellen von Ereignisdatenbanken es gibt. Sowohl RSS-Feeds, statische Ereignisdatenbanken als auch APIs *Application Programming Interfaces* verschiederer Webservices sind Teil der Untersuchung und Evaluation zur Generierung ereignisbezogener Metadaten gewesen.

5. Semantische Verschlagwortung und Metadatenmangement:
   Das übergeordnete Ziel des AB2 liegt im *Information Retrieval*, also dem Finden nutzerdefinierter Informationen in einem Informationssystem. Dazu mussten die aus den zuvor genannten Algorithmen produzieren Metadaten über Schnittstellenformate (JSON, XML, etc.) in eine durchsuchbare Datenbank fließen. Ein komplexer Workflow aus Datenbank, Webtechnologie, Virtualiserungs- und Verteilungstechnolgie wurde im Rahmen des Trecvid-Wettbewerbs entwickelt, anhand dessen einerseits Modelle trainiert, andererseits aber auch Abfragen in über 100.000.000 Metadaten generiert werden. Letztere wurden jeweils auf die *Instance Search Tasks*-relevanten *Queries* angewandt. Ferner wurde ein ebenfalls webbasiertes Metadatenmanagement-Tool entwickelt, dass es erlaubt über textbasierte Ausgaben bestimmter Algorithmen eine *Elastic Search*-Datenbank aufzubauen und über ein Frontend durchsuchbar zu machen –in Echtzeit –auch bei mehreren Zehn- bis Hunderttausend Datensätzen.

6. Parallelisierung und Ausweitung Video:
   Ziel war es, die Generierung und Abfrage von Metadaten zu beschleunigen. In der Generierung von Metadaten wurde der Weg der Verteilung der verschiedenen Algorithmen durch Virtualisierungsansätze gewählt. Ein schnelles Abfragen konnte durch die Zusammenführung in modernen auf Abfragegeschwindigkeit optimierten Datenbanken, wie PostgreSQL und Elastic Search, erreicht werden.

2.3.2 Analyse von Perspektive, Blickwinkel und Pose

Pose und Kamera-Perspektive

Ein Schwerpunkt war die Verallgemeinerung der Anwendbarkeit von Algorithmen (z.B. Posenerkennung) in fotointensiven Szenarien, beispielhaft untersucht am Datensatz *Roter Teppich* (siehe Kapitel 2.2.3). Der zweite Schwerpunkt war die Evaluation von Verfahren zur Analyse menschlicher Posen (und Blickwinkel). In [74] wurde anhand des erstellten Roten-Teppich-Bild-Testsets systematisch untersucht, wie die verschiedenen Posenerkennungsverfahren auf unterschiedliche Perspektiven und Blickwinkel reagieren. Die untersuchten Verfahren beruhen auf *Convolutional Neural Networks* (CNNs) und sind in der Lage den menschlichen Körper in sog. Keypoints zu unterteilen. Diese erscheinen visualisiert als Skelett wie im letzten Zwischenbericht an verschiedenen Beispielen dargestellt wurde. Die räumliche Anordnung der Körper-Merkmalspunkte erlaubt im Grunde eine regelbasierte algorithmische Interpretation von Blickwinkel und Perspektive bei bekannter Kamera-Optik. Um die Generalisierung dieser Verfahren und die Übertragbarkeit auf andere fotointensive Szenarien zu verstehen, hilft es aber nur bedingt sie auf andere Testsets anzuwenden. Entscheidender ist es die Grenzen der Verfahren unter verschiedenen Aufnahmebedingungen zu kennen. Daher wurde entschieden der Evaluation mehr Gewicht

einzuräumen. Tatsache ist, dass die bekannten Verfahren wie von [125, 33] auf großen heterogenen und bereits annotierten Datenbeständen entwickelt wurden. Daher ist die Übertragbarkeit auf andere fotointensive Szenarien wie Fußballspiele, Volksfeste, etc. teilweise gegeben. Was für die Evaluation im Vordergrund stehen sollte, sind die methodenbedingten Auflösungsgrenzen solcher Verfahren mit Blick auf absolute und relative Größe von Personen im Bild, deren Auflösung und deren räumliche Relation zur Kamera (z.B. Drehwinkel oder Rotationswinkel des Körpers oder des Kopfes). Um dies zu untersuchen, spielt ein systematisch strukturiertes Bild-Testset eine große Rolle. Es wurden daher in den 2 Workshop-Publikationen folgende Konzepte verfolgt: (i) die Evaluation der menschlichen Posenabschätzung als Funktion von dessen Orientierung zur Kamera und (ii) die Erstellung physikalisch exakter Bild-Testsets und das Einsparen von Annotationsaufwand zur *Ground Truth*-Erstellung durch ein 3D-Modellierungs-Konzept.

(i) Zur Evaluation wurde das Annotationstool „Deep Learning Web Lab Annotator“ erweitert, um modell-getrieben zeiteffizient annotieren zu können. Dies bedeutet, die vom Posenerkennungsverfahren (Modell) gelieferten Posenpunkte werden bereits mit ihrem jeweiligen Label (z.B. linke Schulter hellgrün in Abbildung 8) angezeigt und müssen nur noch korrigiert werden, falls Sie falsch lokalisiert wurden. Nicht erkannte Posenpunkte können neu hinzugefügt werden, während nicht sichtbare Punkte entfernt oder als nicht sichtbar annotiert werden können. Für das Rote-Teppich-Testset wurde ausschließlich die Person auf dem roten Teppich analysiert. Damit entstanden ca. 1000 Bilder mit bis 18.000 (*Ground Truth*) x,y-Koordinaten aus den Kategorien Füße, Knie, Hüften, Schultern, Ellbogen, Handgelenke, Hals, Augen, Nase, Mund und Ohren, jeweils nach Körperseite unterteilt.

Ein in der Community häufig genutztes Datenset ist unter dem Titel „Microsoft COCO - Common Object in Context“ bekannt [82]. Aus der Webseite geht auch eine Metrik hervor, die es erlaubt einen Keypoint-basierten Überlapp zweier Posen zu berechnen. Hieraus geht jedoch nicht direkt hervor, welche Keypoints die Genauigkeit im Falle einer Fehldetektion reduzieren. Daher wurde eine eigene Evaluationsmetrik eingeführt, die untersucht, ob sich in einem festen Abstand zur *Ground Truth*-Koordinate eine Detektion finden lässt oder nicht. Dies ist in Abbildung 30 in Form grüner Toleranzkreise dargestellt, während Detektionen durch Kreuze wiedergegeben werden. Dadurch ist es möglich falsch und richtig Positive und Negative zu identifizieren, die es erlauben, übliche Evaluationsmetriken wie *Recall*, *Precision* und F1-Maß zu bestimmen, für jedes Körperposenelement separat. Letzteres ist in Abbildung 31 dargestellt und zeigt eindrücklich, dass die Güte der Erkennung der unteren Extremitäten sehr hoch ist, während diese zu den oberen Extremitäten und zum Kopfbereich weiter abnimmt. Es ist anzunehmen, dass dies u.a. an der relativen Größe der Objekte liegt. Die Ohren und Augen sind in manchen Bildern gar nicht erkennbar, werden jedoch weiterhin näherungsweise geschätzt. Das neuere Verfahren von Cao *et al.* [33] ist auf dem

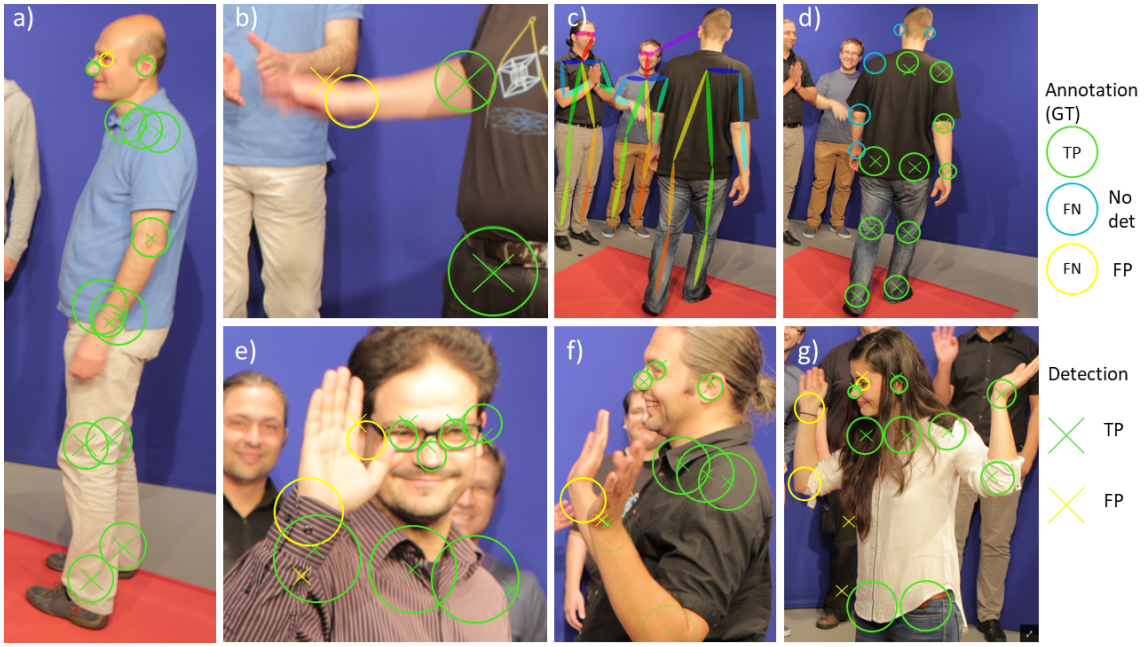


Abbildung 30: Beispieldarstellung des Klassifikationsschemas für menschliche Posen bei verschiedenen Personen, Posen und Blickwinkeln relativ zur Kamera [74].

Testdatensatz auch leicht besser als das Verfahren von Wei *et al.* [125] und erlaubt überdies die Erfassung einiger Gesichtsmerkmale. Das von FlickR randomisiert zusammengestellte Datenset liegt immer unter den Werten des Labor-Datensets. Als Hauptursache ist zu nennen, dass im FlickR-Datenset auch mehrere Personen direkt nebeneinander vorkommen, was zu Fehldetektionen führt, vor allem Falsch Positive, beispielsweise durch Lokalisierung der Extremitäten auf die Nachbarperson und umgekehrt. Dies wird rechts in Abbildung 31 deutlich, wo die FlickR-Detektionen in Einzelpersonen und Gruppen separaiert wurden und sich Ersteres in höherer Genauigkeit ausdrückt, repräsentiert durch das F1-Maß. Der dominante Anteil der Ungenauigkeit der untersuchten Methoden besteht im Auftreten von Falsch-Positiven also niedriger *Precision* bei grundsätzlich hohen *Recall*-Werten.

Während die Kamerahöhe und -winkel im Roten-Teppich-Szenario weitgehend frontal zur Person ausgerichtet waren, erwies es sich im Projektkontext und Abstimmung mit den Stiftern als wichtig, auch Vogelperspektiven zu betrachten. Abbildung 32 zeigt die Aufnahmen einer mit dem S2000 Smart Sensor im Audio-Video-Labor aufgenommen Person beginnend aus der Froschperspektive (Sensornormale = 0°, horizontal auf die Person zeigend) bis zur Vogelperspektive (Sensornormale 90°, vertikal nach unten zeigend). Zur Analyse der Pose wurde der *Open Pose* Algorithmus benutzt [32], in einer Weiterentwicklung aus dem zuvor genannten Abschnitt. Eine wesentliche Erkenntnis der Masterarbeit von Zohair Akhtar mit dem Titel „Evaluation of human pose estimation algorithms from different camera perspectives“ war, dass die Güte mit Annäherung an die Vogelperspektive abnimmt, jedoch bis ca.

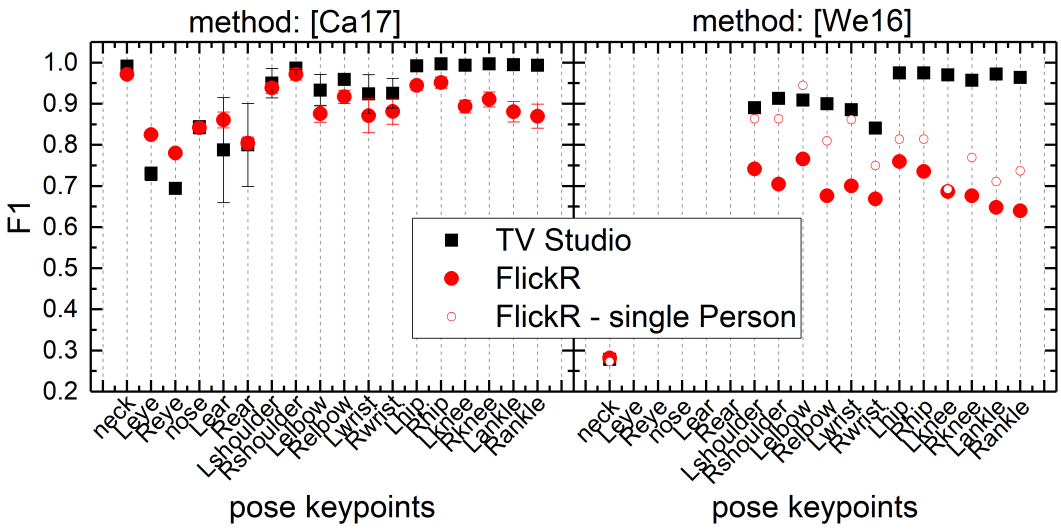


Abbildung 31: Evaluationsergebnisse von menschlichen Posen bei verschiedenen Personen, Posen und Blickwinkeln relativ zur Kamera.

70° nur um 20% [21]. Erwartungsgemäß bricht die Erkennung direkt über der Person zusammen, d.h. verdeckte Körperskelettpunkte werden auch nicht weiter geschätzt, sondern schlichtweg nicht mehr angezeigt. Für viele Kameraszenarien lässt sich aber bis 65° noch gut eine Pose abschätzen. Daher lässt sich ableiten, dass dieser Tote-Winkel-Bereich von 25° unterhalb einer Kamera idealerweise von einem zweiten Sensor abgedeckt werden sollte.

Eine detaillierte Untersuchung des Einflusses der Kameraposition und -orientierung (Perspektive) erfolgte im Rahmen eines Konferenz-Papers bei der HCI International 2019 (HCI - Human Computer Interaction) [86]. In Abbildung 6 ist ein Beispielbild des in Kapitel 2.2.3 beschriebenen Datensatzes zu sehen. Es zeigt eine Person mit Fahrrad, erstellt mit der 3D-Grafiksuite Blender. In der 250 Frames umfassenden Szene läuft die Person mit Fahrrad unter 9 virtuell positionierten Stereo-Sensoren entlang. Diese Szene wurde mit 3 verschieden Abständen zwischen den Sensoren und 5 verschiedenen Neigungswinkeln jeweils als Video gerendert und mit der o.g. Posenerkennung von *Open Pose* ausgewertet. In den Abbildungen 34 zeigt sich, dass die Fehler mit größerem Abstand zwischen 2 Sensoren zunehmen, was auf die geringere absolute Größe oder die schrägeren Perspektiven auf die Person im Bild zurückzuführen sein könnte. Die geringsten Fehler ergeben sich bei einer Detektion nahe an der Person, aber nicht direkt darüber, und deckt sich mit den in Abbildung 33 gefundenen Real-Life-Ergebnissen. Außerdem nehmen die Fehler bei Verdeckung durch das Fahrrad (Richtung Sensor 9) zu. Die Variation der Sensorneigung zwischen -40° und +40° offenbart, dass zwischen -40° bis -20° sehr geringe Fehlerraten nahe Null für die Posenerkennung durch *Open Pose* erzielt werden können (siehe Abbildung 35). Die Fehlerrate nimmt jedoch bei positiven Winkelwerten deutlich zu, bei denen die Person im Wesentlichen von hinten

Abbildung 32: Aufnahmen einer Person unter verschiedenen Sensorwinkeln, variiert in 9 diskreten Schritten zwischen Frosch und Vogelperspektive. Die angegebenen Winkel von 10° bis 170° ergeben halbiert den relativen Winkel zwischen Sensornormale und Horizontale im Raum.

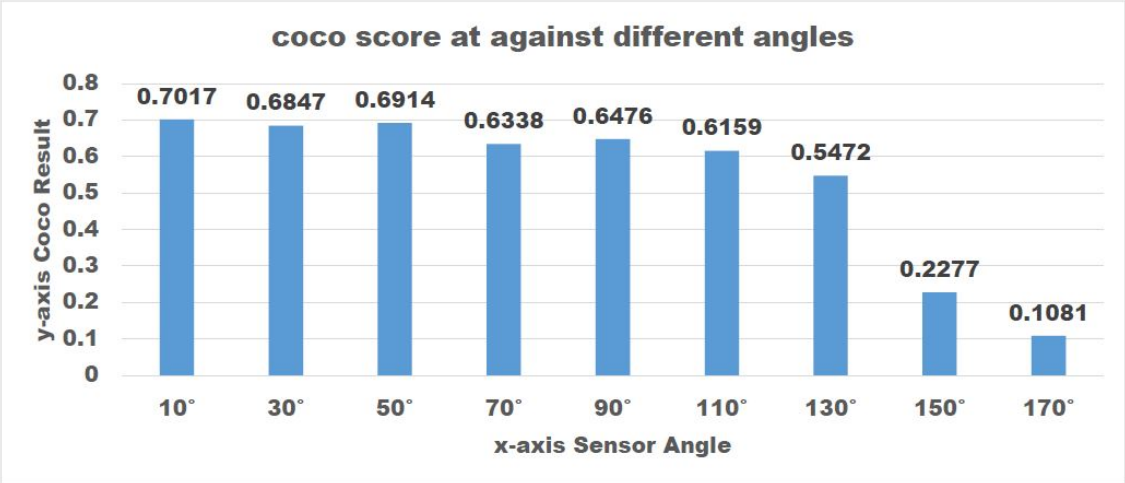


Abbildung 33: Abhängigkeit der Metrik für Posen aus dem Coco-Wettbewerb ([82]) vom Winkel der Kameranormale, vgl. 32. Hinweis: Die angegebenen Winkel entsprechen den 2-fachen Werten der realen Änderung.

erfasst wird. Dies offenbart die Schwäche der Detektion und erlaubt die Schlussfolgerung, dass auch hier zur robusten Ereigniserkennung mehrere Sensoren mit Überlapp und unterschiedlicher Orientierung hilfreich sein können, die es mindestens einem Sensor erlauben Personen von vorn zu erfassen.

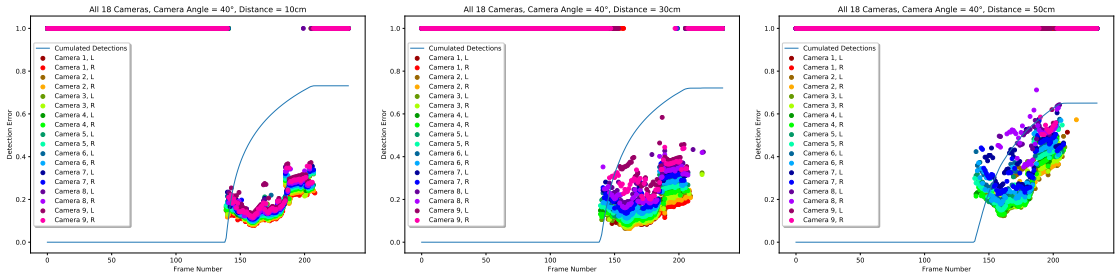


Abbildung 34: Zunahme des Abstandes zwischen den Sensoren bei gleichzeitigem Anstieg der Varianz des Detektionsfehlers und leicht ansteigender Zahl der Gesamtdetektionen.

Die in Kapitel 2.2.8 dargestellten Ergebnisse zur Personenzählung in einem nachgestellten Busszenario zeigen, wie durch die Fusion von 3 Sensoren u.a. Posenerkennungsalgorithmen zur Zählung der Personen (Fahrgäste) im Raum trotz verschiedener Posen und Verdeckungen dennoch eingesetzt werden können. Hier wurde auf die Erfahrung aus den Evaluationen der Winkelabhängigkeit zurückgegriffen und die Sensoren durch leichte Neigung und gezielten Überlapp so positioniert, dass die Fehlerrate nahe Null gedrückt werden konnte. Offen bleibt, wie vergleichbare Szenarien mit vielen Personen auf engem Raum funktionie-

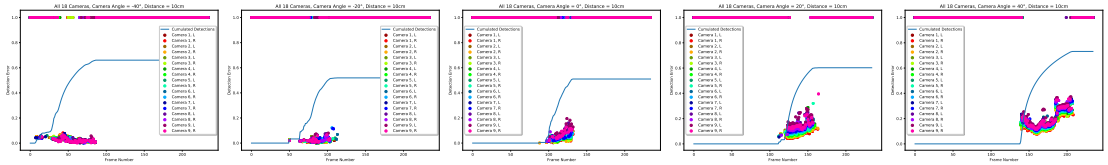


Abbildung 35: Die Variation der Sensorneigung von -40° bis -40° zeigt eine spätere Detektion aufgrund der Bewegung der humanoiden Person, aber auch einen Anstieg der Varianz. Die Gesamtzahl der Detektionen nimmt mit Abnahme in Richtung des vertikalen Winkels ebenfalls ab; die meisten Körperpunkte werden hierbei durch die Person selbst verdeckt.

ren. Eine zum Berichtserstellungszeitraum laufende Masterarbeit beschäftigt sich mit der Frage, inwiefern die direkte Kopferkennung das Personenzählproblem verbessern kann.

Die in Kapitel 2.2.8 vorgestellten Ergebnisse zur FFahrraderkennung aus der Vogelperspektiveßind auch für diesen Arbeitsbereich interessant. Es konnte gezeigt werden, dass die Qualität für Metadaten aus dem Bereich Objekterkennung deutlich gesteigert werden kann, wenn man bestehende Algorithmen auf das konkrete Szenario anpasst. Im Falle der Aufnahme von Fahrrädern von oben, lässt sich so die Erkennungsquote (mAP) vom einstelligen Bereich auf über 80% anheben, in Einzelfällen sogar auf über 90%.

Blickwinkel

In Kooperation mit der Professur Graphische Datenverarbeitung an der Fakultät für Informatik der TU Chemnitz wurde ein neues Bild-Testset-Konzept erstellt, das einen Workflow zur Erleichterung von Annotationen enthält, welches detailliert als Workshop-Paper veröffentlicht wurde [73]. Dazu wurde ein sogenannter Body-Scanner genutzt, der aus 36 zeitgleich aufgenommenen Kamerabildern ein 3D-Modell erstellt. Das räumliche Schema dazu ist in Abbildung 36 zu sehen.

Aus den Einzelaufnahmen, die in Abbildung 37 exemplarisch zu sehen sind, wird ein 3D-Modell des Oberkörpers berechnet. Mittels eines eigens entwickelten 3D-Annotationswerkzeugs wurden 7 Gesichtsmerkmale annotiert: Augen, Ohren, Nase und Mundwinkel. Für jedes dieser Bilder lässt sich dadurch eine Orientierung relativ zu jeder Kamera bestimmen und durch Rotationswinkel Pitch, Roll und Yaw mathematisch beschreiben. So wurden in ca. 30 Minuten 36 Einzelaufnahmen mit jeweils 36 verschiedenen Blickrichtungen aufgenommen. Die 36 erhaltenen 3D-Modelle lassen sich in wenigen Minuten annotieren und aus der 3D-Koordinate, die jeweiligen 2D-Koordinaten im 2D-Bild berechnen. Somit erhält man nicht nur $36 \cdot 36$ Bilder mit je 7 Annotationen, also ca. 9000 Gesichtsmerkmalspunkte in wenigen Minuten, sondern ebenfalls zu den 1296 Bildern jeweils 3 Rotationswinkel (Pitch, Roll, Yaw) in äußerst hoher Genauigkeit. Zudem erhält

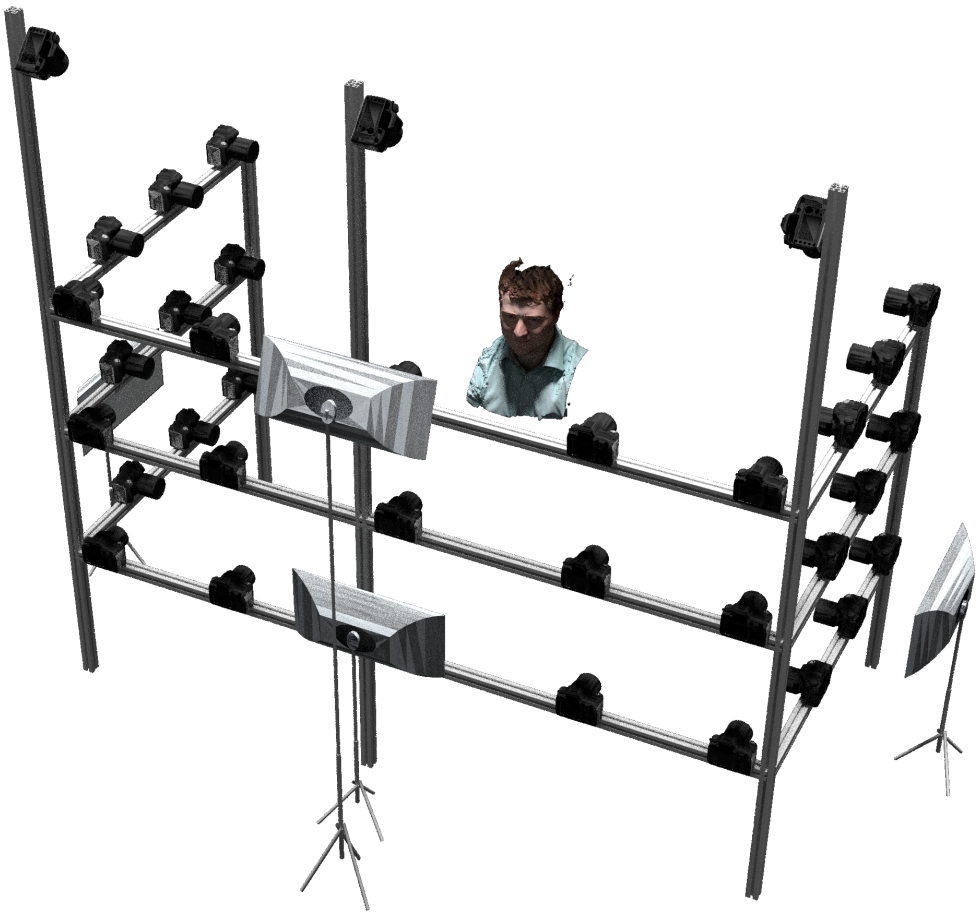


Abbildung 36: Aufbau des aus 36 Kameras bestehenden Body Scanners zur 3D-Modellerstellung des Oberkörpers.

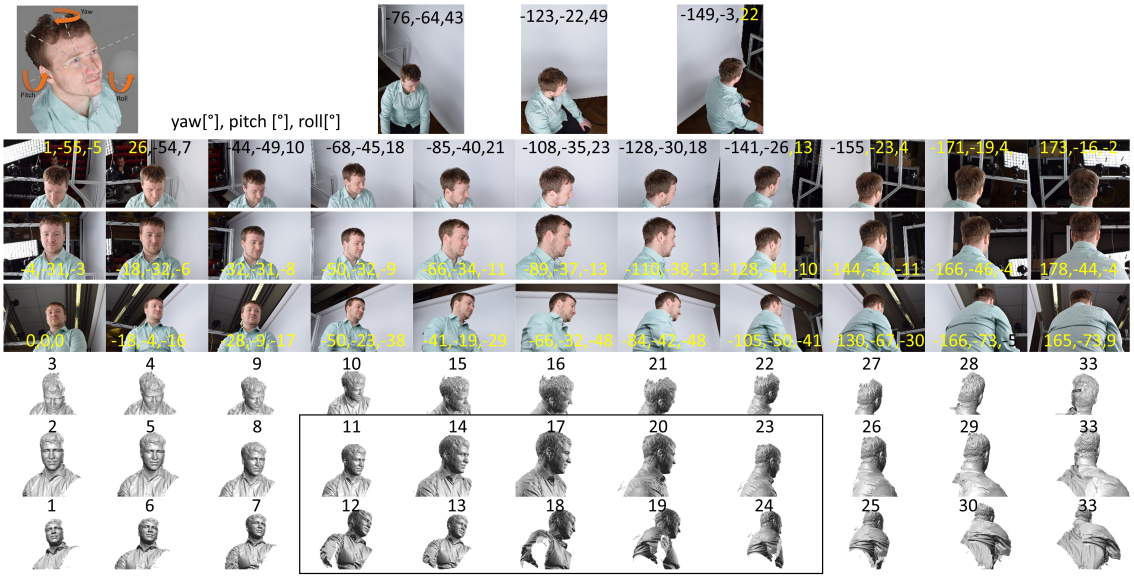


Abbildung 37: Beispielaufnahmen eines zeitlich synchron ausgelösten Schnappschusses der 36 Kameras aus Abbildung 36 mit verschiedenen Kopfposen und den errechneten *Pitch*, *Roll*, *Yaw*-Winkeln des Kopfes relativ zur Kamera.

man für jedes Bild eine Binärmaske des Oberkörperbereichs, der es ermöglicht den Körper (größenskaliert) in diverse Hintergründe einzuarbeiten und damit eine Augmentierung des Datensets um einen beliebig hohen Faktor zu ermöglichen, bei dem Hintergrund und die relative Größe im Bild beliebig variiert werden können. Auf dieser Basis können zukünftig insbesondere Kopfposen-Erkennungsverfahren evaluiert oder trainiert werden. Ferner ist es mit einem Vollkörper-3D-Scanner möglich, den Ansatz auf Ganzkörperaufnahmen zu erweitern. Ein solcher Scanner wurde gegen Projektende in der Professur Graphische Datenverarbeitung aufgebaut und bezieht ca. 100 Kameras ein. Die erstellen Vollkörper-3D-Modelle konnten jedoch bis zum Projektende nicht mehr für weitergehende Analysen verwendet werden.

Kopfposenanalyse

In einer Studie auf einem kleinen Datensatz wurden 3 Kopfposenalgorithmen auf deren Zuverlässigkeit untersucht. Die in Abbildung 38 als Algorithmus 1, 2 und 3 bezeichneten Verfahren beruhen auf trainierten neuronalen Faltungsnetzen (CNNs) [51, 18], aber auch auf klassischen maschinellen Lernalgorithmen wie *Gradient Boosting* und *Ensembles of Regression Trees* [71]. Die Ergebnisse der Evaluation der 3 Algorithmen, dargestellt in Abbildung 38, zeigen die Kopfdrehungen verschiedener Lagewinkel (Roll-Nick-Gier-Winkel, auch bekannt als *roll-pitch-yaw*) im Abstand zur Kamera von 50 cm (links) und 100 cm (rechts). Alle Algorithmen erweisen sich als vergleichbar, weisen jedoch leichte Unterschiede auf. Algorithmus 1 und 2 sind robuster als Algorithmus 3, wenn es um *yaw*-Bewegungen geht, für *roll* und *pitch* ist Algorithmus 2 leicht vorn. Festzustellen ist, dass im Bereich von 45° bis 60° alle

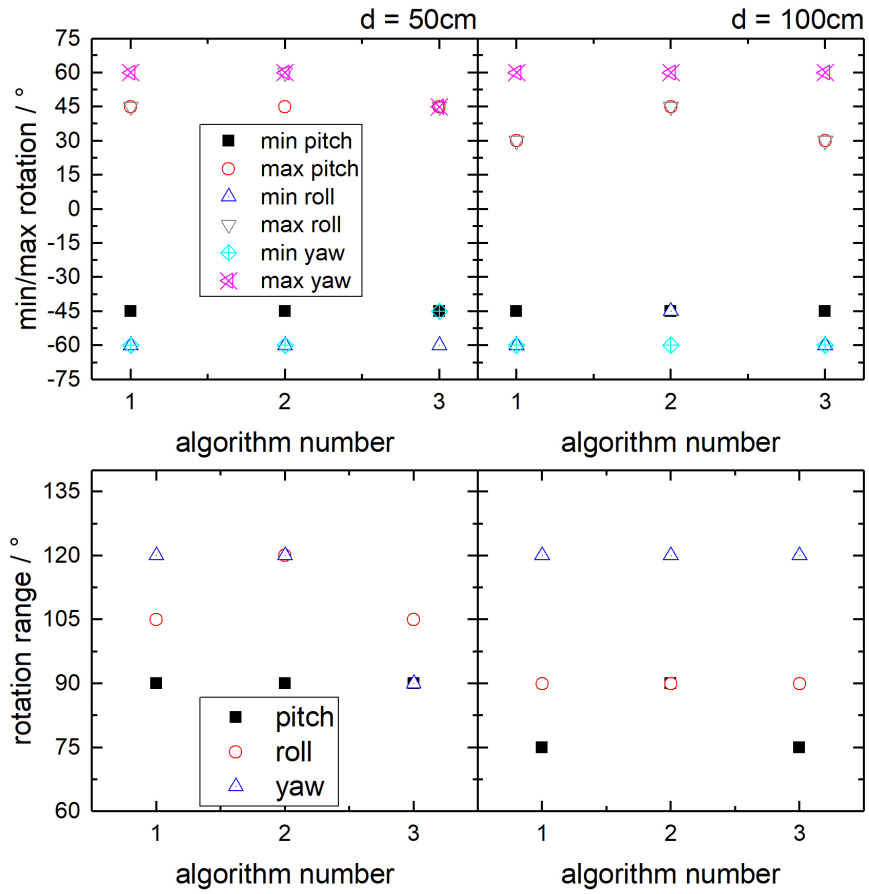


Abbildung 38: The minimum and maximum rotation graphs possible for 3 algorithms with pitch, yaw and roll values for every head rotation at distance 50cm and 100cm was depicted in the first row with an multitude angle. The possible head rotation range graph was depicted for 3 algorithms in the second row.

Algorithmen zunehmend schlechter werden. Zur Metadatengenerierung wäre Algorithmus 2 aus dem Datensatz [93] wohl die beste Wahl.

2.3.3 Verortung von Bilddaten mit GPS mit Ortsnamen

Es wurde ein Matlab-Programm entwickelt, dass die Web-API (Application Programming Interface) von Geonames[7] zum *Reverse Geocoding* (RC) nutzt. Unter Reverse Geocoding versteht man das Gewinnen von Ortsinformationen (wie z.B. Land, Stadt, Straße) aus GPS-Koordinaten. Ein Screenshot der grafischen Nutzeroberfläche dargestellt in Abbildung 39 zeigt die einzelnen Module des Programmes. Es verfügt über folgende Funktionalität:

- Laden und Anzeige eines Bildes aus einem Ordner

[7] `http://www.geonames.org/`

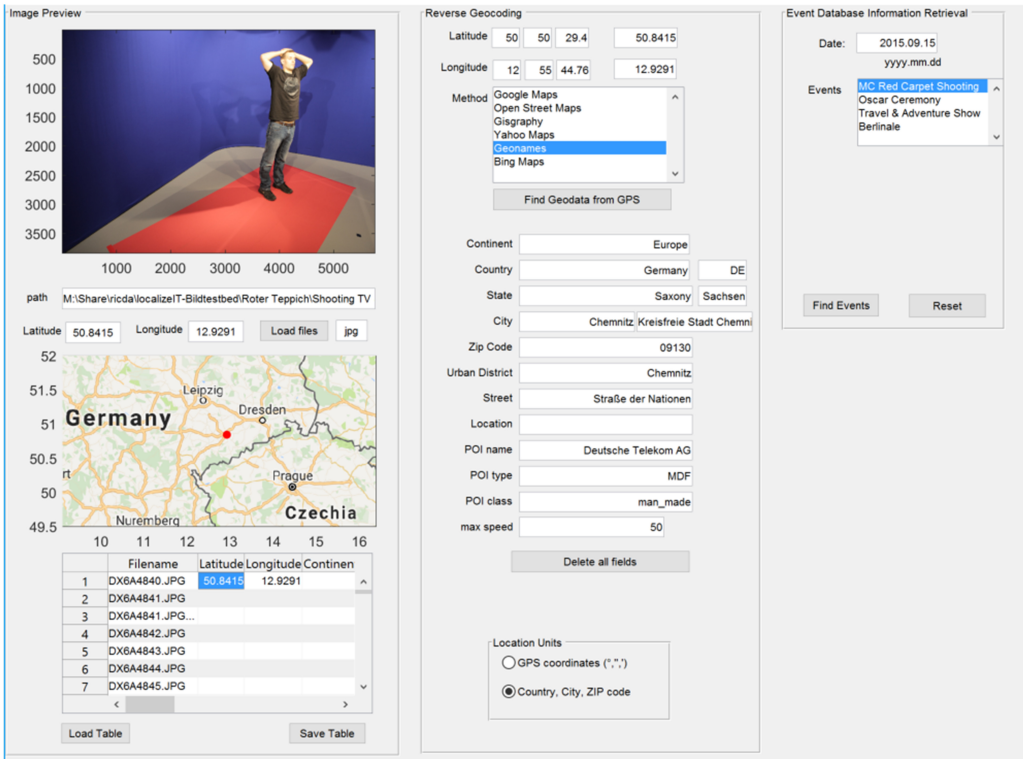


Abbildung 39: Screenshot einer Matlab-basierten Anwendung zur automatisierten Ermittlung von Ortsinformationen aus GPS-Daten (Reverse Geocoding). Das Beispiel zeigt die Ergebnisse der Web-API von Geonames[8].

- Ermitteln der GPS-Geokoordinaten und Visualiserung des Ortes in einer Landkarte
- Listenerfassung von Geodaten aller Bilder eines Ordners mit Lade- und Exportfunktion als Liste in Form einer CSV-Datei
- Reverse Geocoding via Geonames-API zum Abrufen von 14 Ortsmetadateninformationen
- Vorbereitung für die Anbindung weiterer APIs (Google Maps, OpenStreetMap, etc.)
- Vorbereitung für Anbindung von Eventdatenbank-Informationen

Die Anwendung erlaubt die ermittelten Metadaten zugeordnet zu einem Dateinamen zu exportieren und damit auch außerhalb der Anwendung mit den Ortsinformationen arbeiten zu können. Diese Ortsmetadaten können in die Webanwendung mit Ortsschlagwortsuche eingebunden werden, die in Kapitel 2.3.6 näher vorgestellt wird.

Die Qualität der durch Reverse Geocoding gewonnen Ortsinformationen wurde im Rahmen der Masterarbeit von Ronak Gandhi eingehender untersucht [42]. Es wurde ein browserbasiertes Tool entwickelt, welches drei RC-APIs von Google, OpenStreetMap (OSM)

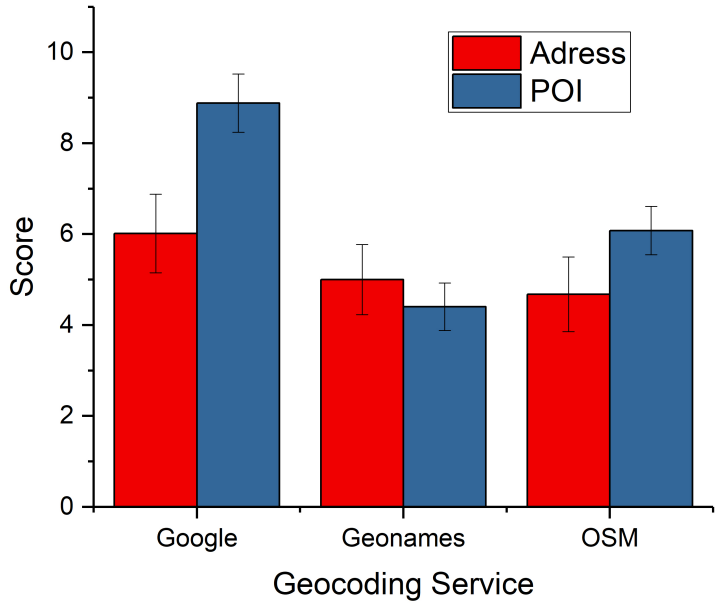


Abbildung 40: Evaluation von Adress- und *Point of Interest* (POI)-Informationen der *Reverse Geocoding*-APIs von Google, Geonames und OpenStreetMap. Für Detailinformationen, siehe Text.

und Geonames einbindet und es erlaubt zu nutzerdefinierten Bildern Geoinformationen der 3 APIs automatisiert abzurufen, diese mit *Ground Truth*-Geoinformationen zu versehen und daraus einen Score zu berechnen. Für AB2 sind die POIs besonders interessant, weil darin konkrete für das *Retrieval* wichtige Suchbegriffe zu erwarten sind, z.B. ein konkreter Veranstaltungsort (PEB-Studentenkeller, Richard Hartmann Halle, etc.). Für Adressen gab es je Bild maximal 7 Punkte zu erreichen für die Informationen: Land, Bundesland, Stadt, Stadt (politisch), Postleitzahl, Straßenname und Hausnummer. Für POIs wurde eine Top-10-Rangliste ausgewertet, wobei es 10 Punkte gab, wenn der gesuchte POI auf Platz 1 war und einen Punkt wenn er auf Platz 10 war. Die Plätze dazwischen wurden entsprechend ihrer Position in 1er-Schritten von 1 bis 10 bewertet. Die Ergebnisse sind in Abbildung 40 zu sehen. Die API von Google stellt sich als beste Anwendung heraus, während Geonames und OSM für Adress-Details vergleichbar sind, jedoch OSM etwas besser für POIs abschneidet. Die erzeugten Metadaten werden im JSON-Format bereitgestellt und sind damit in das Metadatenmanagement-Tool implementierbar, welches in Kapitel 2.3.6 genauer beschrieben wird.

2.3.4 Verortung GPS-loser Bilddaten mit Ortsnamen

Zur Verortung GPS-loser Bilder besteht ein Lösungsansatz darin, diese mit Bildern welche GPS-Daten enthalten zu vergleichen. Daher wurde zunächst der Arbeitsansatz geprüft, inwieweit sich menschlich empfundene Ähnlichkeiten auf numerisch bestimmte Ähnlichkeitsmetriken in Bildern anwenden lassen. In [101] konnte eine Korrelation am Beispiel der FUZZ-Metrik aus ImageMagick [9] bestätigt werden. Das definierte Kriterium für menschliche empfundene Bildähnlichkeit bestand darin, dass sich die Anzahl der sichtbaren Objekte nicht ändern durfte. Es zeigte sich, dass es einen markanten Anteil an Bildern gibt, die von Menschen und durch die FUZZ-Metrik gegensätzlich bewertet werden. Weitere Ähnlichkeitsmetriken aus ImageMagick wurden bereits auf dem in [101] verwendeten Datensatz berechnet. Die Kombination diverser Ähnlichkeitsmetriken birgt das Potential probabilistisch über Ähnlichkeit zu entscheiden. Beispielsweise wurden Bilder mit einem FUZZ-Wert von 0.15 in 50% der Fälle als ähnlich in den anderen 50% als nicht ähnlich eingeschätzt. Nach diesem Prinzip können weitere Metriken genutzt werden um Konfidenzbereiche zu definieren. Der Lösungsansatz von Ritter *et al.* Ähnlichkeitsmetriken in Bildern zu verwenden, geschah auf den Daten des 2016er TrecVid-Wettbewerbes, wobei es galt, bestimmte Charaktere der Fernsehserie East-Enders an bestimmten Orten zu lokalisieren [101]. Die Weiterentwicklung des Sequence Clustering Verfahren aus Markus Rickerts Doktorarbeit [99] führte zu einer Verbesserung der *Accuracy* von $49,7\,\%$ auf $72,8\,\%$ auf den Videodaten des Shot 0. Damit konnte in Szenen auf denen die Konfidenz bei der Personendetektion geringer ist, weil die Person verdeckt oder von hinten zu sehen ist, dieser Konfidenzwert neu skaliert werden, indem das Wissen über die Zugehörigkeit des Bildes zum gleichen Shot genutzt wurde. Das Prinzip ist in Abbildung 8 der 2016er TrecVid-Publikation[67] dargestellt.

2.3.5 Abgleich von Orten und Ereignissen/Metadatenanreichung mit Ereignisdaten

Auf Basis des Arbeitspaketes in welchem eine Übersetzung GPS-basierter Koordinaten in Adressinformationen ergründet wurde (Kapitel 2.3.4), erfolgt hier der Abgleich dieser, aus Bilddaten gewonnenen, Ortsinformationen mit Daten aus Ereignisdatenbanken. Die Problemstellung inkludiert die Erkenntnisse aus Kapitel 2.3.4 und nutzt diese, um Bildern Ereignisse zuzuordnen. Diese Zuordnung basiert neben dem ortsbasierten Abgleich zudem auf einem zeitlichen Abgleich. Hierfür wurde das in der folgenden Abbildung dargestellte Vorgehen entwickelt und umgesetzt.

Für die Umsetzung werden vier Komponenten benötigt. Komponente 1 zum Abruf der Ereignisdatenbanken, Komponente 2 zur Übersetzung der Orts- und Zeitangaben in ein festgelegtes Schema sowie Komponente 3 zum Abgleich dieser Informationen mit den Meta-

[9] `http://www.imagemagick.org`

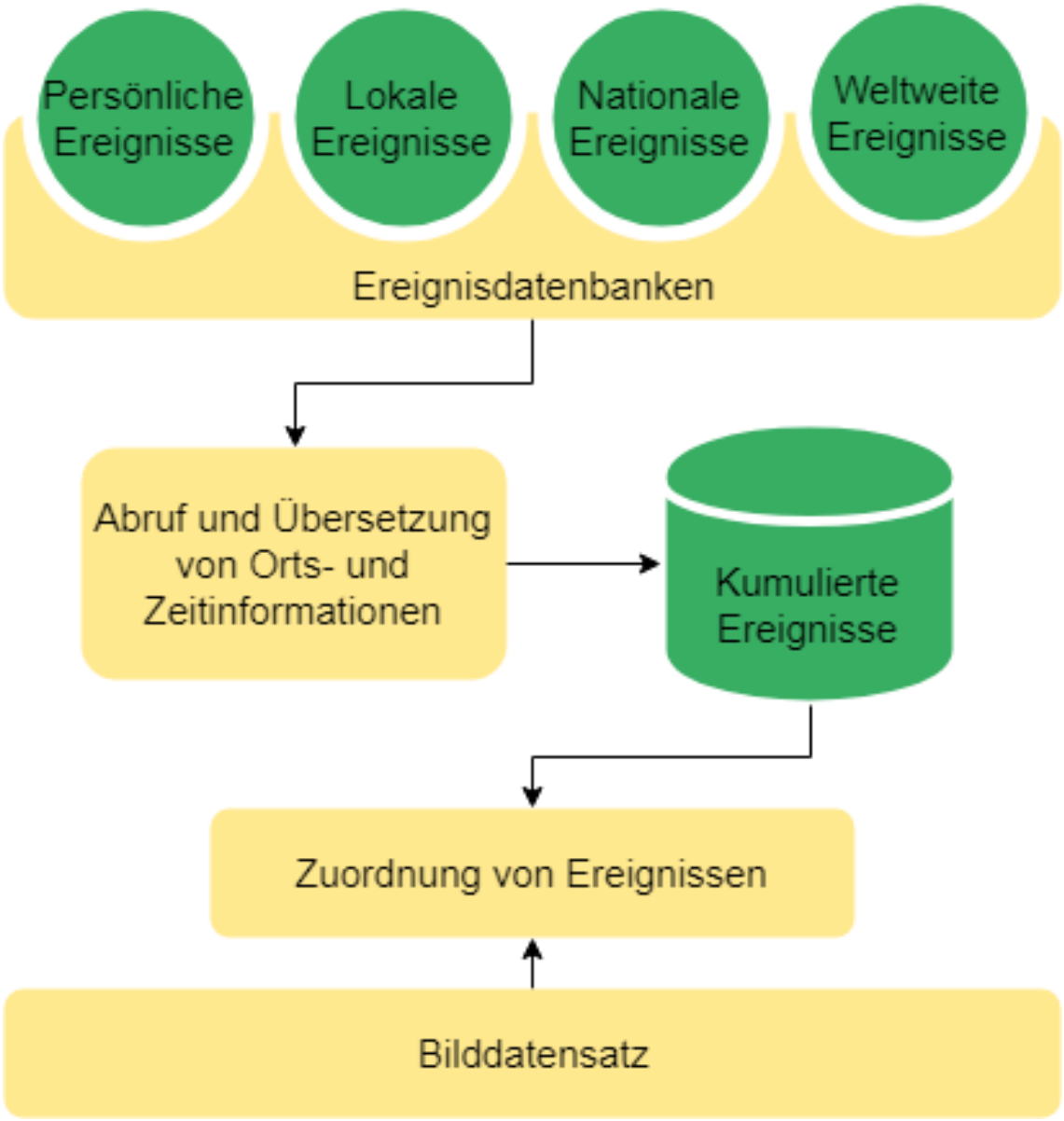


Abbildung 41: Übersicht über die Komponente eines Tools, welches es ermöglicht Ereignisse aus verschiedenen Ereignisdatenbanken zu konsolidieren und für einen Ablgeich mit Bilddaten bereitzustellen. Der Abgleich ordnet einzelnen Bildern Ereignisse auf Basis von Zeit- und Ortsinformationen zu.

informationen des entsprechenden Bildes und der Zuordnung zueinander. Die letzte Komponente bezeichnet eine geeignete Datenbank zur Ablage der Daten.

Neben der strukturellen Analyse erfolgte eine Auswahl an Quellen für Ereignisdaten. Diese sind im folgenden aufgelistet: TU Chemnitz Veranstaltungskalender[10], Wikipedia[11], Eventbrite[12], AllEvents[13]. Für diese Datenquellen erfolgte die Entwicklung entsprechender Adapter, um die Ereignisse in eine konsolidierte Datenbasis zu überführen.

In der Masterarbeit von Raj Tailor wurde das Thema weiter vertieft. Die Analyse auf Vollständigkeit des Veranstaltungskalenders der TU Chemnitz wurde in den folgenden Event-Kategorien in einer Stichprobe von 50 Datensätzen untersucht: Name, Startzeit, Endzeit, Veranstaltungsort und GPS-Koordinate. Während der Name und der Veranstaltungsort immer vorhanden waren, fehlten in ca. 25 % der Fälle die GPS-Informationen und in ca. 8 % der Fälle die Uhrzeiten für Start und Ende. Die Datenbank stellt damit ausreichend genaue Informationen zur Verfügung, um Sie für Event-Metadaten zu nutzen. Die APIs von Eventbrite und Allevents liefern notwendige Eventinformationen bei Abfragen eines bestimmten Datums und einer Stadt. Die Adressdetails 50 gesuchter *Ground Truth*-Adressen wurden mit den *Geocoding* APIs von Google, Allevents und Eventbrite verifiziert, wobei nur 35, 43 und 45 Adressen entsprechend korrekt gefunden wurden. Dennoch stellen beide Plattformen damit eine zuverlässige Event-Datenquelle dar, die zu 90 % zuverlässige Werten liefern sollte. Auch die Abfragezeiten für 100 Events sind mit ca. 45 s und ca. 57 s für Allevent und Eventbrite in einem Rahmen, der es erlauben sollte einige Hunderttaussend Bilder in wenigen Tagen zu verschlagworten. Limitierend sind hier eher die begrenzte Anzahl freier Abfragen pro Tag. Vertiefend wurden verschiedene Abstandsradien zwischen 50 m und 200 m um die Bild-Koordinate herum untersucht auf die Frage hin, ob sich darin der gesuchte Eventort in der erstellen Eventdatenbank befindet. Die Ergebnisse sind in Abbildung 42 dargestellt und zeigen, dass die Richtig-Positiv-Rate den größten Sprung (von unter 50 % auf fast 90 %) zwischen 50 m und 100 m aufweist und auf über 90% ab 150 m anwächst. Ab 150 m Abstand nimmt die *Precision* leicht ab, was für *Information Retrieval*-Aufgaben kein größeres Problem darstellen sollte. In einer möglichen Anfrage an die Bilddatenbank wird i.d.R. eine Top-10 bis Top-100 Liste an Treffern angezeigt und es ist bereits ein Gewinn, wenn darin das gesuchte Bild enthalten ist. Daher würde man für diese Zwecke eine höhere Falsch-Positiv-Rate um wenige % sicher in Kauf nahmen, wenn dadurch der *Recall* entsprechend groß wird, das richtige Event also mit in der Liste ist. Abschließend lässt sich feststellen, dass sowohl spezifische Eventdatenbanken wie des TU Chemnitz Veranstaltungskalenders als auch internationale Anbieter wie Eventbrite und Allevents gute Ressourcen darstellen,

[10] `https://www.tu-chemnitz.de/tu/termine/`
[11] beispielsweise `https://de.wikipedia.org/wiki/Oscar`
[12] `https://www.eventbrite.de`
[13] `https://allevents.in`

um Eventinformationen abzurufen. Die APIs von Wikipedia und Google wurden aus Zeitgründen nicht eingehender untersucht, sind grundsätzlich aber für den Einsatz zur Gewinnung spezifischer Eventinformationen ebenfalls denkbar. Für das im Antrag erwähnte Rote-Teppich-Szenario wären z.B. alle Oskarverleihungen in Wikipedia mit Datumsinformationen zu finden und für die ganze Historie dieser Veranstaltung via API abfragbar.

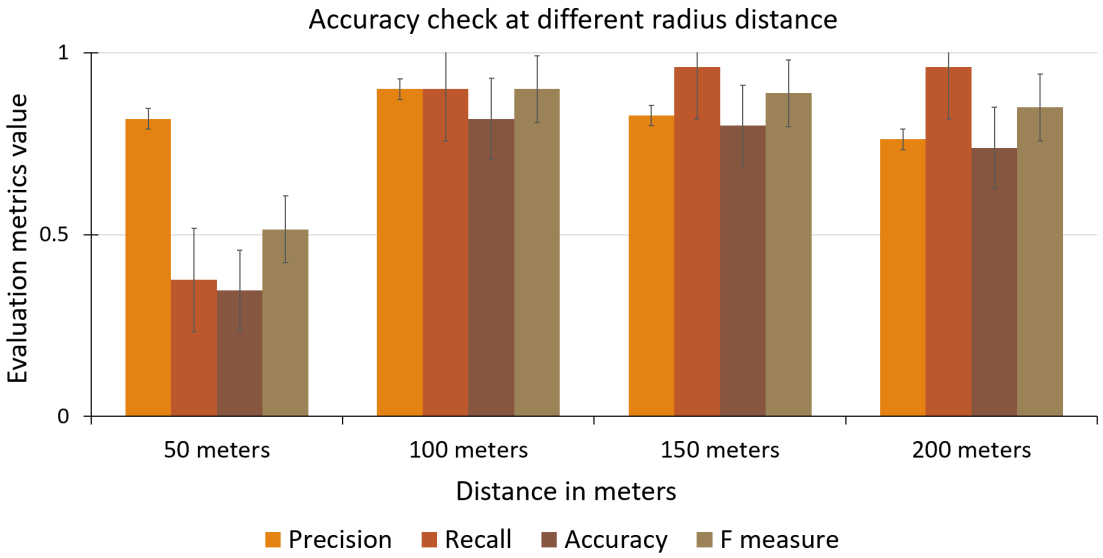


Abbildung 42: *Accuracy* check at different radius distance

2.3.6 Semantische Verschlagwortung und Metadatenmangement

Metadaten-Typen und State-of-the-Art Algorithmen

Im Rahmen des Forum Bildverarbeitung 2016 in Karlsruhe konnte in einem Vortrag mit dem Titel „Detektion und Klassifikation von Personen anhand demografischer Merkmale durch Convolutional Neural Networks“ gezeigt werden, wie CNNs, welche auf einem Teilkorpus des ImageNet-Datensets trainiert wurden, sehr gute Klassifizierungsergebnisse in den Kategorien Personen-Detektion, Alter, ethnische Zugehörigkeit und Geschlecht liefern können.

Im Rahmen des wissenschaftlichen Projektworkshops bei den Chemnitzer Linux-Tagen 2019 wurde ein Artikel verfasst, der nochmals die essentiellen Bild-Metadaten-Algorithmen vorstellt, die für diesen Arbeitsbereich relevant sind [75]. Im Rahmen des TRECVID-Wettbewerbs wurden die damit erzeuten Metadaten in eine PostgreSQL-Datenbank überführt. Technische Details befinden sich in den Publikationen von Roschke und Thomanek *et al.* [124, 123]. Die folgenden Metadaten-Algorithmen sind im Beitrag [75] genau referenziert:

- Skelett-Körperpunkte/Posen –Openpose

- Objekterkennung –Detectron
- Objekterkennung –Yolo9000
- Orts-/Szenenerkennung –Places365
- Szenenbeschreibung –Show and Tell
- Farbextraktion –Colorthief
- Gesichterkennung –FaceRecognition

Ergänzt werden können diese um die in den vorangegangenen Kapiteln vorstellten Algorithmen:

- Kopfposenerkennung, Kapitel 2.3.2
- Geoinformation - Kapitel 2.3.4
- Eventinformationen - Kapitel 2.3.5

Metadaten-Management-Plattformen

Entwicklung eines holistischen verteilten maschinellen Lern-Frameworks für den TRECVID-Wettbewerb

Wie bereits in Kapitel 2.2.7 erwähnt, entstand in den Jahren 2018 und 2019 ein umfangreiches auf Webtechnologien basierendes maschinelles Lern-Framework (nach dem Vorbild von AMOPA, [100]), das Metadaten aus manuellen Annotationen verwaltet, aber solche auch aus den o.g. Algorithmen und Objekterkennungs-Frameworks generiert, dabei jedoch auf die Anforderungen des Wettbewerbs zugeschnitten ist [123]. Typische Abfragen an das System, z.B. Personen an einem bestimmten Ort zu finden, ist hier zwar eingeschränkt auf Personen und Orte, die es im Wettbewerb gibt, wäre jedoch grundsätzlich erweiterbar um generelle Abfragen. Bemerkenswert ist, dass es trotz weit über 100.000.000 Datenbankeinträgen möglich ist, innerhalb von nur ca. 100 ms alle für den Wettbewerb relevanten Abfragen als Tabelle oder über eine Visualisierungsschnittstelle geliefert zu bekommen. Abbildung 43 zeigt die 2018 erzielten Ergebnisse. Diese liegen zwar hinter der Konkurrenz, zeigen jedoch, dass der gewählte algorithmische Ansatz und das Retrieval wettbewerbsfähig sind. Abschließend muss betont werden, dass dies nur durch die kooperative Zusammenarbeit mit Prof. Ritter (Hochschule Mittweida) möglich war, der das Projekt in den ersten beiden von insgesamt fünf Jahren geleitet hat. Insbesondere der Wissens- und Technologietransfer des 2016 in LocalizeIT betreuten Masterabsolventen und anschließend von Prof. Ritter betreuten Doktoranden Christian Roschke zusammen mit Robert Manthey (AB 1) führten zu gemeinsamen erfolgreichen Einreichungen bei dem

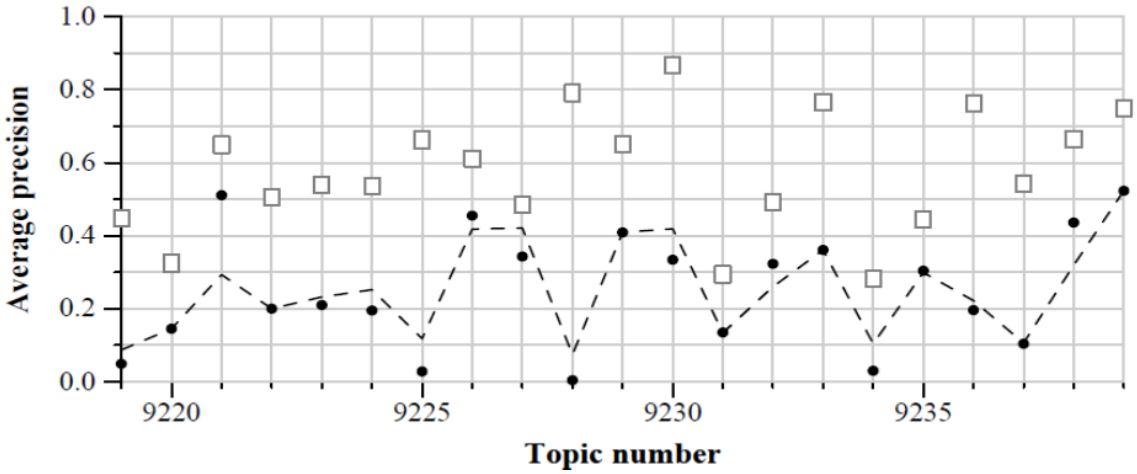


Abbildung 43: Ergebnisse im sog. *Interactive Run* der TRECVID-Evaluationskampagne in der Kategorie „INS – Instance Search“. Offene Quadrate zeigen die besten Ergebnisse im Wettbewerb, geschlossene Kreise unsere erzielten Ergebnisse. Die Aufgaben sind verschiedene Darsteller an bestimmten Orten zu finden. Details, siehe Kapitel 2.2.7.

international bekannten TRECVID-Wettbewerb. Mit Blick auf die Autorenliste zeigt sich die fruchtvolle Zusammenarbeit von Doktoranden und Postdocs der Medieninformatiker (einschließlich Media Computing) der TU Chemnitz und der Hochschule Mittweida [124, 123, 121].

Entwicklung eines Bildmetadaten-Retrieval Tools nach dem Stockphoto-Webseiten-Prinzip

Um den im Vergleich zum TRECVID-Wettbewerb allgemeineren *Retrieval*-Anforderungen dieses Arbeitsbereichs gerecht zu werden, wurde ein weiteres webbasiertes Tool nach dem Vorbild von Bildsuch-Webseiten entwickelt, auch bekannt als Stockphoto-Webseiten, wie z.B. Gettyimages ([14] und Adobe Stockphotos ([15]. Eine Vorarbeit und einen Überblick über den *State of the Art* dazu stellte die Masterarbeit von Rafay Ahmed dar [20]. Die Weiterentwicklung zu einem funktionalen Bildmetadaten-Retrieval-Tool mit grafischer Nutzeroberfläche (*Frontend*) ist in Abbildung 44 zu sehen. Die Anwendung enthält einen generischen Filtermechanismus, der es erlaubt, mehrere Abfragen mit logischen UND-Verknüpfungen zu kombinieren. Die Freitexteingabe in das Eingabefeld (*Field*) durchsucht alle für die Freitextsuche markierten *Keys*. Nachteilig kann sein, dass die Keys bekannt sein müssen. Neben der Suche nach Schlagworten (*Keys*) kann aber auch der Wertebereich gefiltert werden *Values*. Im Beispiel werden Bilder gezeigt, die eine Bank, jedoch nicht

[14] `www.gettyimages.de`
[15] `https://stock.adobe.com/de/`

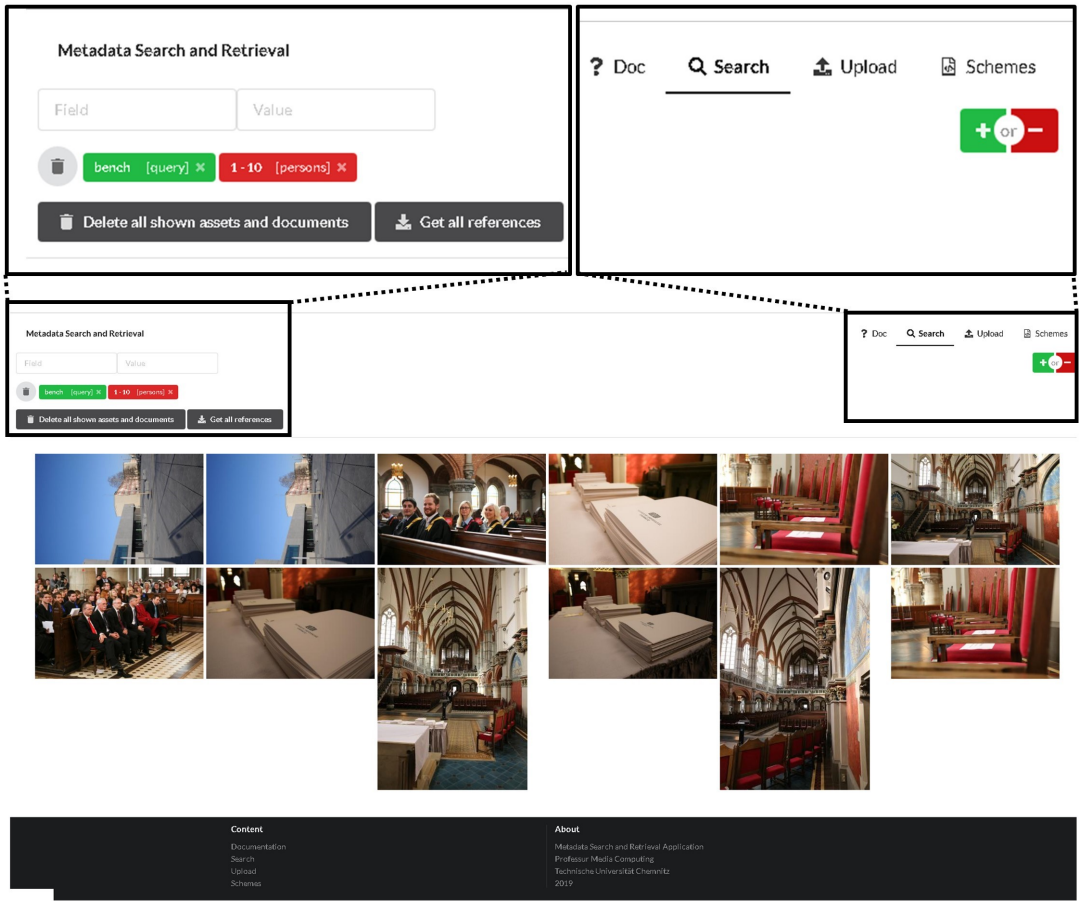


Abbildung 44: Grafische Nutzeroberfläche des webbasierten Bildmetadaten-Retrieval-Tools und Treffer zu Elastic Search-basierter Abfrage von Bildern mit Suchworten „Bank“ und Suchwert „1 - 10 Personen“ aus [75].

1 - 10 Personen enthalten. Die roten und grünen *Buttons* erlauben Positiv- und Negativ-Suchkriterien zu kombinieren. Das gezeigte Beispiel ist aus dem Datensatz der Universitätskommunikation (Beschreibung, siehe Kapitel 2.2.3). Eine umfassende technologische Evaluierung war im Projektzeitraum nicht mehr möglich. Die Arbeit ist demzufolge eher als *Proof-of-Concept* zu betrachten. Die Beschreibung des webtechnologischen Hintergrunds (Graphql API als Schnittstelle zur Metadatengenerierung bzw. -speicherung, für *Retrieval* und Key-Value-basierte Suche über die Nutzeroberfläche, *Elastic Search* zur schnellen indexbasierten Metadatenverwaltung) ist detailliert im zu den Chemnitzer Linux-Tagen 2019 erstellten Workshop-Band zusammengefasst [75].

2.3.7 Parallelisierung und Ausweitung Video

Aufbauend auf den den Erfahrungen der vorangegangenen TRECVid-Teilnahmen und vor allem durch die während des Forschungsaufenthaltes von Robert Manthey 2017 am NIST

in Gaithersburg, USA, bekannt gewordenen und intensiver untersuchte Lösungen zur Detektion und teilweisen Identifikation von Gesichtern, Personen, Objekten sowie deren Verhaltensweisen, wurde die im Berichtszeitraum erstellte und verwendete Lösung in erheblichem Umfang auf Virtualisierungslösungen aufgebaut, wie in Abbildung 45 dargestellt. Die zugehörige, während des TRECVid-Workshops 2017 [84] dem Fachpublikum vorgestellte und auf Docker aufsetzende Vorgehensweise zur Virtualisierung, ermöglicht die parallele Ausführung der Analysesysteme und kann somit die notwendige Verarbeitungsgeschwindigkeit der im Projekt verwendeten Detektionssysteme bereitstellen und darüber hinaus die Entwicklung und Erprobung neuer Systeme und Verfahren erheblich beschleunigen.

Die beim TRECVid-Workshop 2018 von der Forschungscommunity vorgestellten Systeme setzten ebenfalls in erheblichem Umfang auf mit Docker virtualisierte Lösungen, was die Angemessenheit dieser Vorgehensweise bestätigt. So wurden beispielsweise die Objekt-, Gesichts und Personenerkennungs-Frameworks (z.B. Detectron, Yolo und OpenFace) eingebunden, deren Outputs quasi simultan in eine zentrale Datenbank fließen. Hervorzuheben ist die Ausnutzung der GPU, welche die Docker-Virtualisierung technisch gewährleistet. Viele der genannten Algorithmen basieren auf CNNs und werden in ihrer Verarbeitung per GPU entsprechend unterstützt. Verglichen zu CPU-basierter Verarbeitung stellt das eine Beschleunigung von ca. einer Größenordnung in einer Standard-Consumer-Rechner-Konfiguration dar. Für die verschiedenen Anforderungen an Umgebung, Framework und Bibliotheken bietet Docker eine gekapselten Umgebung, die es erlaubt die Algorithmen oder die Verarbeitung gleicher Datenströme auf verschiedene Rechner zu verteilen und Konflikte zu vermeiden (z.B. Version einer Bibliothek). Damit ist eine parallelisierte Verarbeitung großer Datenmengen möglich und war daher sehr gut auf Videos anzuwenden, wie in Kapitel 2.2.7 dargestellt und beispielhaft in Abbildung 15 zu sehen ist.

2.3.8 Bilanz nach Erreichen des letzten Meilensteins in AB 2

In der semantischen Verknüpfung von Bildern gibt es dank des Fortschrittes in der *Computer Vision* Gemeinschaft im Kontext von *Deep Learning* Algorithmen einen großen Sprung. Vor allem in den letzten beiden TRECVID-Wettbewerben konnte erfolgreich gezeigt werden, dass Algorithmen implementiert, angepasst und verfeinert wurden, die ein großes Spektrum bei der Erzeugung von Metadaten abdecken durch: Detektion/Erkennung von Alter, Geschlecht, Hautfarbe von Personen, diversen Objekten aus mehreren Hundert Klassen, ortsbezogener Klassen (Restaurant, Umkleidekabine, Fußballfeld), GPS-basierten Geodaten (Land, Stadt, Straße, Point of Interests), Ereignissen und Tracking einfacher Verhaltensmuster von Personen. Metadaten dieser Art wurden auf Dateiebene strukturiert, z.B. im JSON-Format. Letzteres wiederum kann sehr gut in Suchmaschinen eingepflegt werden, die für Information Retrieval optimiert sind (sog. NoSQL-Programme wie Elastic-

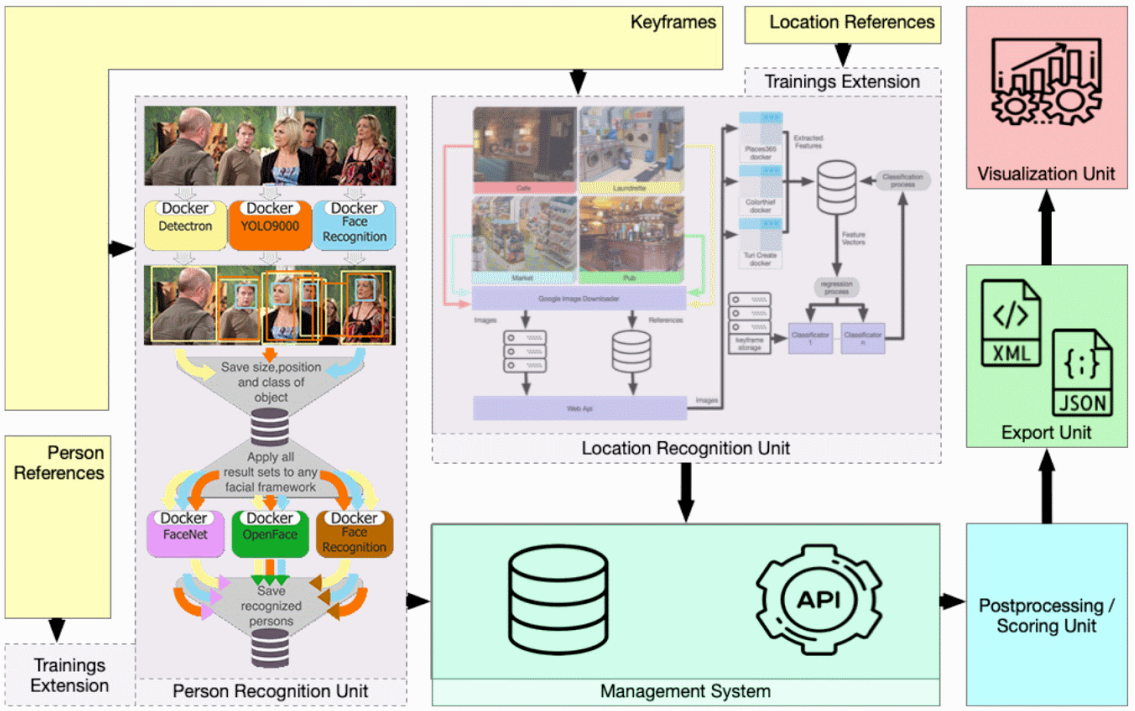


Abbildung 45: Schema des Workflows für den INS-Aufgabenbereich mit den in der Virtualisierungsumgebung Docker parallelisierten Analyse- und Identifikationssystemen, der zentralen Datenspeicherung sowie der webbasierten Management- und Ergebnisdarstellung [121].

Search aber auch objektrelationale Datenbankmanagementsysteme wie PostgreSQL). Im Projektzeitraum wurden weit über 100.000.000 Bildmetadaten generiert und integriert und im Rahmen der TRECVID-Wettbewerbe erfolgreich evaluiert. Weiterhin wurde ein über ein Browser-Frontend bedienbares Tool zu Demozwecken entwickelt, das zusätzlich noch Orts- und Eventdaten integrieren kann. Orts- und Eventdaten-Retrieval wurden auf kleinen Testdatensätzen evaluiert und über JSON-Schnittstellen in die o.g. Systeme integrierbar gemacht. Damit lässt sich illustrieren wie das Suchen in großen heterogenen Datenbeständen zeiteffizient und kontextbasiert möglich ist.

2.4 AB 3: Integration von Audioereignissen in die Echtzeitanalyse

In bestimmten Situationen ist eine videobasierte Lokalisierung unzureichend. Ist beispielsweise zu wenig oder gar kein Licht vorhanden, oder treten unerwartet Verdeckungen durch Objekte auf oder beschlägt auch nur die Linse, können keine brauchbaren Informationen mehr extrahiert werden. Hier kann die Audioanalyse helfen die Lokalisierung im Raum zu unterstützen. Ferner kann eine Audioanalyse helfen, Situationen genauer einzuschätzen, wenn beispielsweise bei der Raumüberwachung überprüft wird, ob eine Person schreit, spricht oder schweigt. Solche Informationen können beispielsweise in der Betreuung von Demenzkranken im eigenen Heim wichtig sein.

2.4.1 Aufgabenstellung und Zielsetzung

Im Projekt wurden Aufgaben quartalsweise formuliert und auf diese Weise auch in den Zwischenberichten abgehandelt. In der folgenden Auflistung sollen jedoch eher die übergeordneten Inhalte betrachtet und deren Lösung kurz angerissen werden:

1. Audiobasierte Sprecher- und Geräuscherkennung:
   Nach dem Aufbau eines Katalogs und der Erstellung entsprechender Testsets sollten Algorithmen zur automatischen Erkennung von Audiosignalen, inbesondere aus dem Bereich des assistierten Lebens (AAL, *Ambient Assisted Living*), entwickelt werden. Dies beinhaltet die Erkennung von Sprache, Nicht-Sprache und diverser Geräusche des alltäglichen Lebens daheim (Haustiere, Geräte, usw.).

2. Szenen- und Verhaltensanalyse durch Audio-Video-Fusion:
   Als Ergänzung zu AB 1 sollen neben getrennten auch gemischte Video- und Audiodaten erzeugt und in die Verhaltensontologie von AB 1 integriert werden. Ziel ist es, in Situationen in denen Videoanalyse an ihre Grenzen stößt (z.B. Verdeckung, wenig Licht, kleine Bildgröße), Audiosignale zu verarbeiten, mit dem Ziel das Szenenverständnis zu verbessern.

3. Beschleunigung von Algorithmen und Nutzung von DSPs:
   Ursprünglich war es angedacht, Audioereignisse und Echtzeitanalyse umfassend auf DSPs zu implementieren und durch den Vergleich mit konventionellen Rechnerarchitekturen zu evaluieren. Dies wurde in geringerem Umfang auch auf klassischen DSPs umgesetzt, jedoch in Absprache mit Stiftern auf FPGA, ARM und GPU erweitert, nicht zuletzt vor dem Hintergrund Algorithmen, die auf *Deep Learning* basieren auf verschiedenen Hardware-Plattformen umzusetzen und auf deren Performanz zu untersuchen.

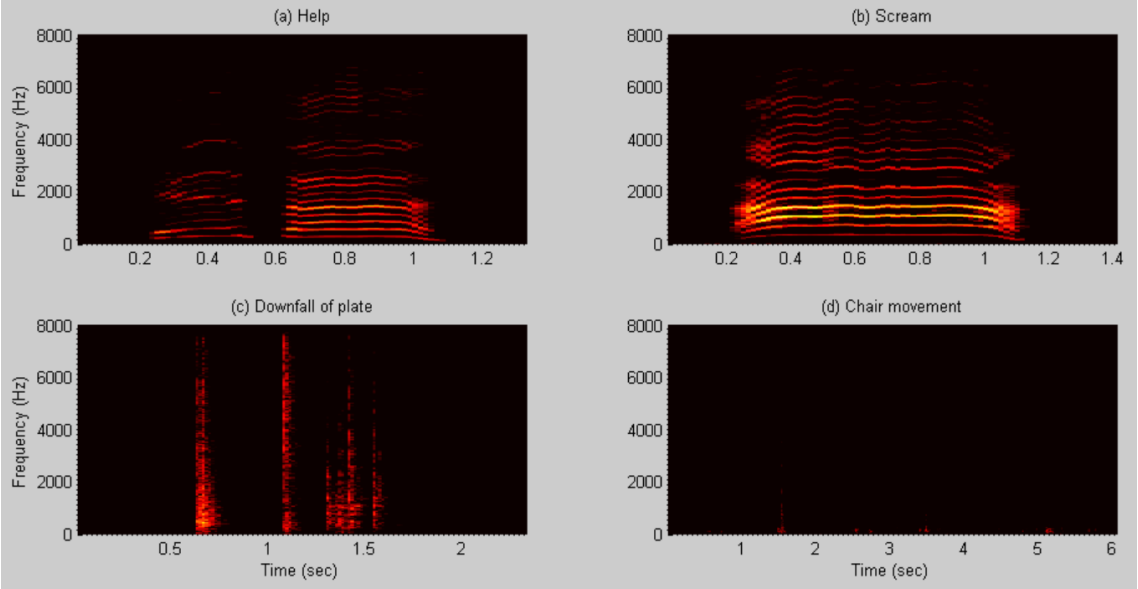


Abbildung 46: Spektrogramm ausgewählter Audioereignisse.

4. Audiobasierte Lokalisation durch Laufzeitunterschiede:
   Die Lokalisierung von Objekten bzw. Personen im Raum durch Verarbeitung von Audiosignalen war ein weiteres Ziel dieses Arbeitsbereiches. Neben der unterschiedlichen Signalstärke kann auch die unterschiedliche Laufzeit des Signals von der Quelle zum Mikrofon genutzt werden, um räumliches Hören zu ermöglichen. Zu nutzen waren dabei die Mikrofon-Arrays des Audio-Video-Labors. Die wesentliche Aufgabe bestand darin, die Lokalisation von Geräuschquellen durch die Analyse von Signalstärke und Laufzeitunterschieden des Audiosignals möglich zu machen.

2.4.2 Audiobasierte Sprecher- und Geräuscherkennung

Vorarbeiten der Audioklassifikation

Das erste Audio-Testbed wurde im Kapitel 2.2.3 unter dem Titel „10 Klassen *Ambient Assisted Living*“ vorgestellt.

Die Analyse der Spektralstruktur der Audioereignisse ist sehr wichtig für die Auswahl der geeigneten Merkmale zur Erkennung dieser Audioereignisse. Abbildung 46 zeigt das Spektrogramm für einige Audioereignisse. Die Spektralstruktur der nicht-sprachlichen Ereignisse und sprachlichen Ereignisse sind unterschiedlich, wobei als Geräusche nicht-sprachliche Ereignisse berücksichtigt werden. Verschiedene Merkmale im Zeit- und Frequenz-Bereich wurden extrahiert, z.B. Energie, Nulldurchgangsrate (Zero-Crossing Rate –ZCR), Grundfrequenz (F_0), Spektrum, Mel Frequency Cepstral Coefficients (MFCCs) und Linear Prediction Coefficients (LPC). Die akustischen Signale wurden mit einer Fensterlänge von 30 ms und

Verschiebung von 10 ms segmentiert. Die Merkmale wurden mit dem Tool „openSMILE“ extrahiert. Neben den originalen Merkmalen (Low-Level Descriptors –LLD) wurden entsprechende *Delta Coefficients* (∆LLD) und *Statistical Functionals* (min, max, range, standard deviation und mean) extrahiert. Folgende verschiedene Teilmengen an Merkmalsvektoren wurden verwendet:

- A (120 Merkmale): MFCCs
- B (50 Merkmale): Energy, pitch, ZCR
- C (230 Merkmale): Energy, pitch, ZCR, spektrale Merkmale
- D (350 Merkmale): Energy, pitch, ZCR, spektrale Merkmale, MFCCs
- E (430 Merkmale): Energy, pitch, ZCR, spektrale Merkmale, MFCCs, LSP
- F (6373 Merkmale): Baseline feature set of the Computational Paralinguistics Challenge (ComParE) in Interspeech 2013

Zur Klassifizierung und Modellerstellung wurden die Daten in Trainingsdaten (80 %) und Testdaten (20 %) geteilt. Die Leistung der Merkmalsvektoren wurde mit mehreren Klassifikatoren untersucht, um die besten Merkmale und Klassifikatoren für die Klassifizierung der akustischen Ereignisse zu bestimmen. Die folgenden *Machine Learning*-Algorithmen wurden mithilfe des „WEKA“ *Data Mining Toolkits* angewendet:

- ClassificationViaRegression (CVR)
- K-Nearest Neighborhood (KNN)
- Sequential Minimal Optimisation (SMO)

Die Ergebnisse der Klassifizierung der akustischen Ereignisse sind in Abbildung 47 dargestellt. Die Performanz wurde mittels F-Measure gemessen, welche das harmonische Mittel zwischen *Precision* und *Recall* ist. Die besten Ergebnisse wurden mit dem Merkmalsvektor (E) und SMO-Klassifikator erzielt [56].

Die Arbeiten mit klassischen Audio-Deskriptoren in Kombinationen mit klassischen maschinellen Lernverfahren wurden von Robert Herms unterstützt, der kurzzeitig im Projekt mitgewirkt hat und innerhalb der Projektlaufzeit seine Dissertation mit dem Titel „Effective Speech Features for Cognitive Load Assessment: Classification and Regression“ finalisierte [55]. In der Arbeit wurde gezeigt, dass kognitive Belastung von Menschen in gewissem Maße durch Sprachanalyse messbar ist.

Vogelidentifikation anhand des Gesangs

In Weiterentwicklung der o.g. eher klassischen maschinellen Lernverfahren, angewandt auf eine Klassifikationsaufgabe mit 10-Audio-Klassen [56], zeigten Forschungsarbeiten von

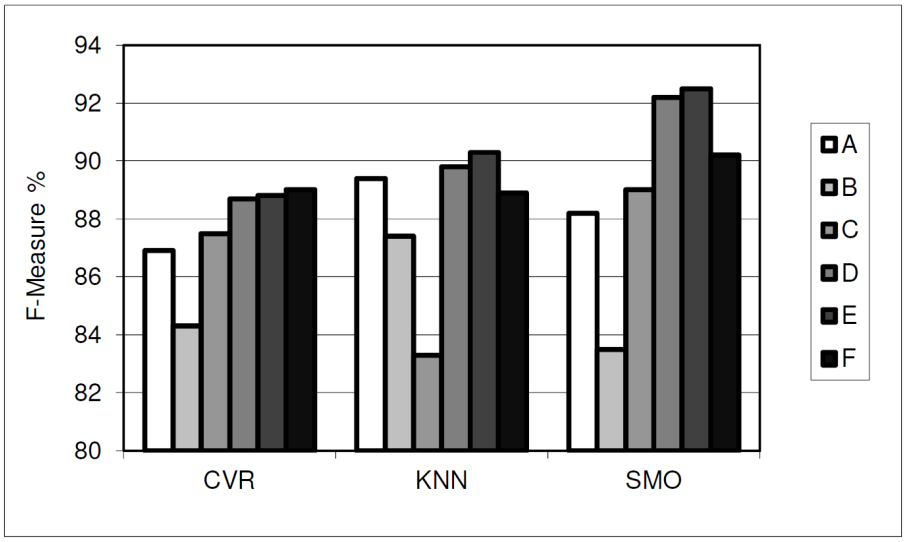


Abbildung 47: Ergebnisse der Klassifizierung.

Stefan Kahl im Projektjahr 2017, dass künstliche neuronale Netze nicht nur in der Objekt- und Personenerkennung sehr gute Ergebnisse liefern [66], sondern auf solch einem eingeschränkten und unter Laborbedingungen erstellen Datenkorpus zu nahezu perfekter Klassifikationsgenauigkeit führen können [65]. In den folgenden beiden Abschnitten wurden diese Ansätze deutlich erweitert auf realitätsnahe Innen- und Außenaufnahmen aus den Forschungsdomänen der Vogelgesangserkennung und der Klassifikation von Geräuschen im *Ambient Assisted Living*-Kontext.

Beim Forschungsaufenthalt im Jahr 2017 von Stefan Kahl bei Prof. Dr. Holger Klinck, dem technischen Direktor am Cornell Lab of Ornithology, Ithaca, New York, wurde der Kontakt vertieft und mit ersten gemeinsamen wissenschaftlichen Ergebnissen im Bereich der Klassifikation von Vogelgesängen untermauert.

Zum Hintergrund: Spektrogramme spielen in der Vogelforschung eine bedeutende Rolle. Wir können davon ausgehen, dass visuelle Darstellungen von Vogelgeräuschen wertvolle Informationen über die Identität der Arten enthalten. Audio-Spektrogramme erwiesen sich als besonders geeignete Darstellung. Tiefe neuronale Faltungs-Netze (CNNs) haben die traditionellen Klassifikatoren im Bereich der visuellen Erkennung und in der akustischen Ereignis-Klassifizierung übertroffen. Dennoch erfordern tiefe neuronale Netze Expertenwissen, um leistungsstarke Modelle zu entwerfen, zu trainieren und zu testen. Mit dieser Beschränkung und den Anforderungen zukünftiger Anwendungen im Auge, wurde eine umfassende Forschungsplattform für die automatisierte Überwachung der Aktivitäten von Vögeln im Rahmen der Doktorarbeit von Stefan Kahl mit dem Titel „Identifying Birds by Sound: Large-scale Acoustic Event Recognition for Avian Activity Monitoring“ entwi-

ckelt [64]. Die Menge an Trainings-, Validierungs-, und Testdaten umfasst mehr als 3.900 Stunden Audio-Aufnahmen, 987 Klassen und fast 300 Stunden vollständig kommentierte Vogel-Klanglandschaften (*Soundscapes*) mit fast 80.000 Vokalisationen. Das Benchmark-System liefert Spitzen-Ergebnisse in verschiedenen akustischen Bereichen. Es wurde sowohl als Expertenwerkzeug also auch für öffentliche Demonstratoren genutzt. Das Framework ist flexibel anpassbar und der wissenschaftlichen Community über Code-Repositories ([63, 62]) zugänglich und damit auch für die künftige Verwendung im Rahmen von Naturschutzmaßnahmen verfügbar. Die Verwendeten Konzepte der Mustererkennung sind robust übertragbar auf unterschiedliche Anwendungsdomänen (siehe folgender Abschnitt, AAL94). Sie lieferten im Rahmen zahlreicher Veröffentlichung *State-of-the-Art* Ergebnisse [68, 70, 69].

Besonders hervorzuheben ist der zweite Platz im BirdCLEF-Wettbewerb 2017 [47, 60], der sowohl in der regionalen als auch in der überregionalen Presse aufgenommen wurde. Details dazu befinden sich im Presseartikel[16]. Hohe Sichtbarkeit konnte zudem durch die Beteiligung als Organisatoren des BirdCLEF-Wettbewerbs 2018 [45] und die dazugehörige Bereitstellung eines Baseline-Systems [70] erzielt werden. Ferner wurde der Wettbewerbsgemeinschaft ein Übersichtsartikel für Wettbewerbsteilnehmer unter dem Titel „Species Prediction based on Environmental Variables using Machine Learning Techniques“ zur Verfügung gestellt [119]. Der Konferenzbeitrag mit dem Titel „Large-Scale Bird Sound Classification using Convolutional Neural Networks“ ist die mit über 30 Zitaten am meisten zitierte Publikation des Projekts, unter den Beiträgen, die sich direkt mit der Vorstellung eines entwickelten Algorithmus beschäftigen [68].

Klassifikation von Geräuschen aus dem *Ambient Assisted Living*-Kontext

Um die auditive Erkennung zu erweitern, wurde der bisher 10 Geräuschsklassen umfassende Korpus erheblich vergrößert. Dabei wurden akustische Szenarien in den Mittelpunkt gestellt, die für die Stifterfirma Intenta von besonderer Bedeutung sind: Umgebungsgeräusche, Geräusche im Kontext von *Ambient Assisted Living* und Geräusche in öffentlichen Verkehrsmitteln.

Die Grundlage für den erweiterten Korpus bildeten neben den eigens aufgenommenen Daten einschlägige, annotierte Audio-Datensets, wie das ESC-50 Datenset [95], das Google AudioSet [43] und das Ultrasound-8K Datenset [106]. Für den Anwendungsbereich Umgebungsgeräusche wurde das ESC-50 Datenset unverändert zur Evaluation verwendet (50 Klassen, 2.000 Audio-Aufnahmen gesamt). Für *Ambient Assisted Living* haben wir ein kombiniertes Datenset aus allen oben genannten Ursprungsdatensets erzeugt. Dieses umfasst 94 Klassen und eine Gesamtmenge von 18.884 Audio-Aufnahmen. Für den Bereich Geräu-

[16] `https://localize-it.de/2018/01/15/spitzenplatz-bei-wettbewerb-zur-klassifikation-von-vogelgesaengen/`

sche in öffentlichen Verkehrsmitteln haben wir 1.994 Aufnahmen aus dem Google AudioSet verwendet und in einem Datenset mit 2 Klassen zusammengefasst.

Die Motivation für derart große Datensätze entstammt dem Ergebnis, dass wir in der Vergangenheit sehr gute Klassifikationsraten mit den künstlichen neuronalen Netzarchitekturen Stefan Kahls im Bereich der Vogelstimmenklassifikation erzielen konnten [68], [69], [70]. Diese künstlichen neuronalen Netze (hier: CNNs) benötigen jedoch eine sehr große Menge an Trainings-Daten, weshalb wir versucht haben einen möglichst großen und diversen Korpus zu generieren.

Die Grundlage für den eigentlichen Geräusch-Klassifikator bildete Stefan Kahls Architektur zur Vogelstimmenklassifikation [68]. Der vorliegende Quellcode [63], [62] diente als Ausgangspunkt für weitere Anpassung an die vorliegende Problemstellung und die eingangs genannten Anwendungskontexte.

Das initiale Setup der Software-Frameworks zum Trainieren und Evaluieren des Klassifikators gestaltete sich wie zur erwarten kleinteilig und war mit entsprechendem Zeitaufwand verbunden. Auch die verschiedenen Experimente und die Vielzahl an iterativen Verfeinerungen der Klassifikator-Architektur waren sehr zeitaufwändig. Insbesondere die Trainings-Zeiten, die beim Trainieren unseres Netzes, und im Allgemeinen bei künstlichen neuronalen Netzen, notwendig sind, machten sich hier besonders bemerkbar.

Dieser Zeitaufwand hat sich jedoch bezahlt gemacht. Unser Klassifikator ist in der Lage Klassifikationsraten zu liefern, die über den Ergebnissen der bisherigen Klassifikatoren der wissenschaftlichen Community für das ESC-50 Datenset liegen. Wir führen mit 88,75 % *Mean Accuracy* damit momentan die ESC-50 Benchmarkliste [94] an. Wir hoffen nach der wissenschaftlichen Publikation im Rahmen unseres Workshop-Tagungsbandes bei den Chemnitzer Linux-Tagen 2019 [109] in die Benchmarkliste aufgenommen zu werden. Unsere eigenen Klassifikations-Ergebnisse aus [65] konnten wir damit deutlich verbessern. Abbildung 48 zeigt die Ergebnisse unseres Klassifikators im Vergleich mit unserer Baseline und Ergebnissen aus der Literatur. Erwähnenswert ist, dass auch eine Sprechererkennung enthalten ist, die zumindest männliche und weibliche Sprecher unterscheiden kann. Die sog. *Speaker Diarization* ist die Detektion der Anfangs- und Endzeit oder der Segmente, die zu jedem Sprecher gehören. Sie wird hier zumindest in einer zeitlichen Auflösung von 2s bis 5s gewährleistet. In der nachgelagerten Phase könnte nun die Sprecher-Identifikation stattfinden, wobei unbekannte Sprecher identifiziert werden können. Nach der Recherche der verwendeten Tools zur Sprechererkennung werden die folgenden Tools verwendet. Das LIUM_SpkDiarization Tool wird zur *Speaker Diarization* verwendet [89, 104] und das ALIZE Tool zur Sprecher Identifikation [29].

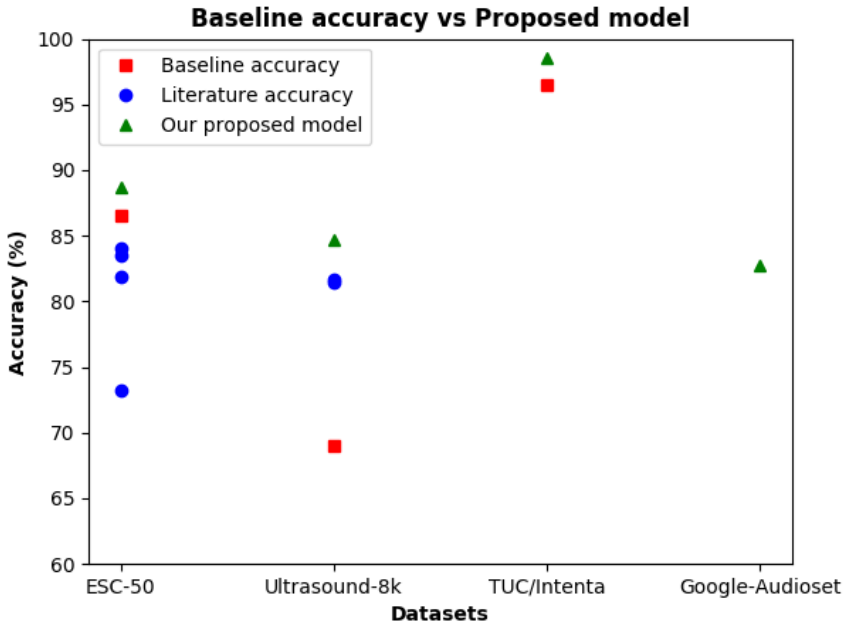


Abbildung 48: Vergleich der Klassifikations-Genauigkeiten für verschiedene Datensätze: Unsere Baseline vs. unser vorgeschlagenes Modell vs. andere Literatur.

Für das kombinierte Datenset bestehend aus 94 Klassen erreichten wir eine Validierungs-Genauigkeit von 70,66 %. Das Datenset zu Geräuschen des öffentlichen Verkehrs (2 Klassen) konnte mit einer Validierungs-Genauigkeit von 92,25 % klassifiziert werden. Der Klassifikator des 94-Klassen-Datensets wurde im Rahmen einer Publikation zu einem entwickelten Demonstrator auf der international renommierten Konferenz ACM Multimedia 2019 veröffentlicht [107]. Der Klassifikator wird im Folgenden als AAL94-Klassifikator bezeichnet, bezugnehmend auf die 94 detektierbaren Klassen aus dem Bereich *Ambient Assisted Living*.

Zusammenfassend lässt sich festhalten, dass wir mit unserer Forschung und unseren Ergebnissen im Bereich des *State-of-the-Art* und leicht darüber hinaus rangieren.

Annotationswerkzeug für Audio-Analyse und -Material

Für die Analyse von Audiodaten wurde ein webbasiertes Annoatationstool entwickelt, das in Kapitel 2.2.4 unter „Audio, Bild, Videoannotationsframework von Christian Roschke“ bereits kurz vorgestellt wurde und im Folgenden noch etwas näher erklärt wird: Das Tool ist webbasiert und der Server besteht aus einem Hauptsystem und mehreren beliebig hinzufügbaren Plugins (Abbildung 49). Das Hauptsystem beinhaltet die komplette Management-Logik. Dabei werden unter anderem Methoden bereitgestellt, um Nutzer zu verwalten, Applikationsbereiche zu schützen, Views zu generieren und Plugins zu integrieren. Die Verwendung von Plugins ermöglicht die einfache Erweiterung des funktionalen Kerns, ohne

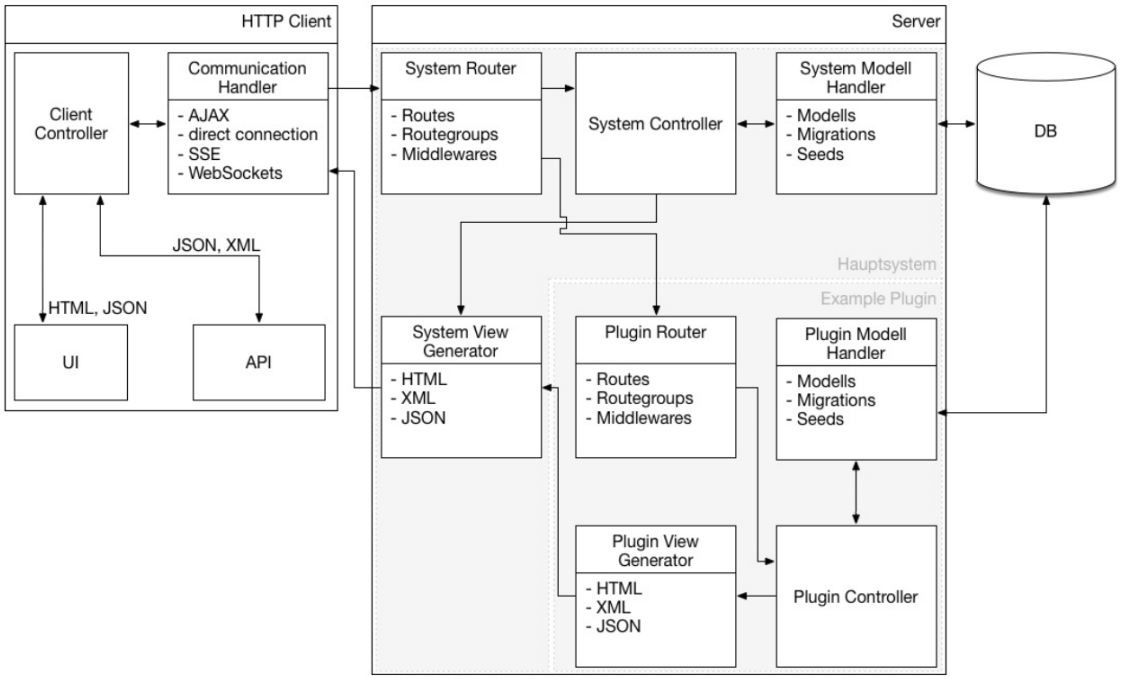


Abbildung 49: Schematische Darstellung der Architektur für das Management-System des Annotationstools.

das Risiko die Kernlogik zu gefährden. Der System-View-Generator nutzt die im Controller erstellten Datenstrukturen und wandelt diese in ein standardisiertes Austauschformat, wie beispielswiese HTML, JSON, oder XML um. Die erzeugten Formate werden dann als Antwort an den Communication-Handler des HTTP-Client übertragen.

Neben der Variante Anfragen an den System-Controller zu übergeben, ist es auch möglich, diese an ein Plugin weiterzureichen. Jedes in das System integrierte Plugin besteht wie das Hauptsystem aus Router, Modell-Handler, View-Generator und Controller. Die im Plugin-Controller mit Hilfe des Plugin-Modell-Handler erstellten Datenstrukturen werden bearbeitet und daraufhin in ein beliebiges Austauschformat gewandelt und in den View des System-View-Generator integriert.

2.4.3 Szenen- und Verhaltensanalyse durch Audio-Video-Fusion

Im Rahmen des TRECVID-Wettbewerbs 2019 (allgemeine Details, siehe Kapitel 2.2.7) wurde der AAL94-Klassifikator aus dem vorherigen Abschnitt neu trainiert. Dazu wurde ein eigenes Annotationstool entwickelt, dessen Beschreibung Kapitel 2.2.3 unter „Videobasiertes Audioklassen-Label-Tool“ zu entnehmen ist. Die damit erstellten ca. 1.900 Trainingsdaten aus über 20 Audioklassen wurden mit den bereits vorhandenen ca. 19.000 Audiodateien neu trainiert und auf den über 450 Stunden umfassenden TRECVID-Korpus angewandt.

Die erzeugten Audio-Metadaten wurden in die Datenbank integriert, wo auch sämtliche Bild- und Videometadaten verwaltet wurden (siehe Abbildung 45). Die erzielten AAL94-Klassifikationsergebnisse wurden im „Run 1“ verwendet und mit den anderen Metadaten (Bild/Video) in einem Punktesystem gewichtet. Ferner wurden die im Unterauftrag an Prof. Ritter erwähnten *Boosting*-Verfahren angewandt. Die Ergebnisse in diesem Jahr lagen jedoch durchweg bei nahe 0 %, was den Median-Ergebnissen zufolge bei vielen Teilnehmern der Fall war und vermutlich auf die gestiegenen Anforderungen der zu lösenden Aufgaben zurückzuführen ist. Ein Vergleich des Systems, ohne AAL94, wurde bislang nicht vorgenommen. Daher lässt sich abschließend keine klare Aussage treffen, inwiefern Audioklassifikation und videobasierte Metadaten zu einer besseren Szenenbeschreibung genutzt werden können.

2.4.4 Beschleunigung von Algorithmen und Nutzung von DSPs

Aufgrund ihrer enorm hohen Leistungsfähigkeit, ist es neben klassischen DSPs auch attraktiv, GPUs und FPGAs in der Audiosignalverarbeitung einzusetzen. Eine Herausforderung stellt hierbei jedoch die Komplexität in der Programmierung dieser beiden Technologien dar. Daher werden DSP-Systeme, mit ihrer wesentlich einfacheren Handhabung und hohen Flexibilität, weiterhin von Bedeutung sein. Zusammen mit den Stiftern wurde das Konzept erarbeitet, dass auch FPGAs, ARM-Prozessoren und GPUs als DSP-Plattformen aufzufassen sind, und im Vergleich mit konventionellen Rechnerarchitekturen evaluiert werden sollten. Daher haben wir diese gemeinsam definierten DSP-Systeme in Form von Benchmarks gegeneinander verglichen. Im Vordergrund stand hierbei die Leistungsfähigkeit der individuellen Systeme hinsichtlich ihrer Verarbeitungsgeschwindigkeit bei Aufgaben der Audio-Merkmalsextraktion sowie -Klassifikation. Es wurde empirisch, quantitativ untersucht, wo Stärken und Schwächen der ausgewählten Systeme liegen und deren Echtzeitfähigkeit bewertet. Abbildung 50a zeigt die 4 zum Benchmarkvergleich ausgewählten DSP-Plattformen und deren jeweilige Schlüsseleigenschaften.

Für die Benchmarks wurde auf allen Systemen das Framework „openSMILE“ zur Audio-Merkmalsextraktion und das auf maschinelle Lernverfahren spezialisierte Klassifikations-Framework „Weka“ eingesetzt. Primär wurden die Ausführungszeiten für verschiedene Teilaufgaben des Klassifikationsprozesses evaluiert. Diese umfassen einerseits die Merkmalsextraktion sowie andererseits die Klassifikation. Insgesamt wurden 6 verschiedene Merkmalssätze und 5 verschiedene Klassifikations-Algorithmen evaluiert. Dabei fand als zu klassifizierender Datensatz ein eigens erstellter Korpus mit 10 bis 20 unterschiedlichen Geräuschklassen aus dem Bereich des *Ambient Assisted Living* Anwendung. Insgesamt wurden 140 Einzelexperimente durchgeführt.

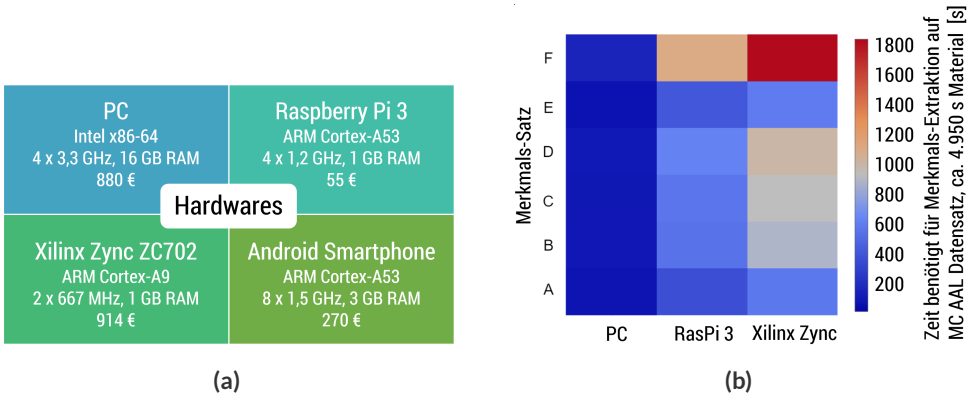


Abbildung 50: a) Schlüsseleigenschaften der ausgewählten DSP-Systeme im Vergleich. **b)** Ausführungszeiten bei der Merkmalsextraktion für Linux PC, Raspberry Pi 3 und Xilinx Zynq ZC702.

Abbildung 50b zeigt einen repräsentativen Teil-Ausschnitt aus der Ergebnismenge der Experimente. Die Grafik stellt eine Heatmap der benötigten Ausführungszeiten bei der Extraktion verschiedener Mermalssätze auf verschiedenen Plattformen dar. Die einzelnen Merkmalssätze sind mit den Buchstaben A bis F kodiert. Es wird deutlich, dass keine Beschleunigung durch die ausgewählten eingebetteten DSP-Systeme im Vergleich zum PC stattfindet. Desweiteren wird ersichtlich, dass erhebliche Unterschiede in der Rechenzeit verschiedener Merkmalssätze bestehen. Dies liegt in der jeweiligen Größe der Merkmalssätze und deren individuellen, rechnerischen Komplexität begründet.

Betrachtet man die Gesamtheit der Experimente, wird deutlich, dass auch hinsichtlich der benötigten Rechenzeit der verschiedenen Klassifikatoren erhebliche Unterschiede bestehen. Abbildung 51 zeigt das Gesamtranking der untersuchten Plattformen.

Wie wir bereits vor den Tests erwartet hatten, bewerkstelligt der Linux PC die Klassifikation insgesamt am schnellsten, da dessen CPU mit 3,3 GHz Taktfrequenz und 4 Kernen die größte Raw-Processing-Power aufweist. Knapp dahinter reiht sich das Android-Gerät ein, welches mit 8 Kernen zu je 1,5 GHz ebenfalls über eine hohe Rechenleistung verfügt. Der Raspberry Pi 3 und das Xilinx Zynq-Board sind gegenüber den vorgenannten Plattformen weit abgeschlagen und belegen absteigend nach Prozessor-Taktfrequenz die letzten beiden Plätze im Ranking.

Zusammenfassend lässt sich sagen, dass das Android-Gerät dem Linux PC als Baseline-Technologie hinsichtlich der Verarbeitungsgeschwindigkeit in den untersuchten Aufgabenkategorien sehr nahe kommt. Dies ist besonders interessant, da Android-Geräte im Allgemeinen mobil sind, einen erheblich geringen Energiebedarf haben als PCs und im vorliegenden Fall, das Gerät nur ca. ein Drittel der Anschaffungskosten im Vergleich zum PC

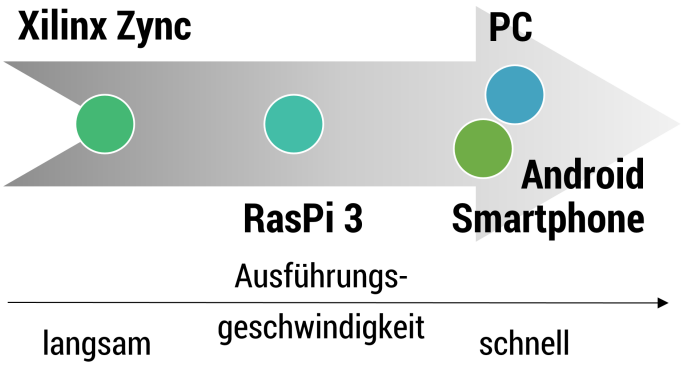


Abbildung 51: Gesamt-Ranking der verschiedenen DSP-Plattformen hinsichtlich ihrer Ausführungsgeschwindigkeiten bei Aufgaben der akustischen Merkmalsextraktion und -klassifikation.

aufweist. Zudem sind Android-Geräte sehr weit verbreitet. Ein weiteres interessantes Resultat ist, dass der Raspberry Pi 3 gegebenenfalls echtzeitfähig im hier betrachteten Anwendungskontext ist. Besonders bedeutsam ist dies, da der Raspberry Pi 3 ein enorm preisgünstiges System ist. Das Xilinx Zynq-Board ist gemessen an der Verwendung dessen ARM-Prozessors nicht empfehlenswert im hier betrachteten Kontext, da es die teuerste aller getesteten Plattformen und gleichzeitig aber die leistungsschwächste darstellt. Im Kontext dieses Arbeitspaketes konnte eine Masterarbeit betreut und erfolgreich zum Abschluss gebracht werden [120].

Die State-of-the-Art Schlüsseltechnologien zur Bearbeitung rechenintensiver Aufgaben in eingebetteten, mobilen Systemen umfassen heute: CPU/DSP, GPU und FPGA. Zwei von uns betreute Studierende konnten diese Technologien im Zeitraum 2017/2018 im Rahmen ihrer Abschlussarbeiten nutzen und jeweils eindrucksvoll zur Anwendung bringen. Carolin Dürrling entwickelte ein System zur Detektion und Lokalisation von Folgetonhörnern auf CPU/DSP [39], während Martin Dörfelt eine CNN-basierte Personenerkennung auf FPGA implementierte [38].

Auch der Export der trainierten Modelle zur Vogelstimmenklassifikation (siehe Kapitel 2.4.2) ist flexibel und unterstützt eine Reihe von Plattformen. Dazu zählen Workstations, ARM-Architekturen, Smartphones (BirdNET-APP) oder Webanwendungen[17]. Die verwendeten Konzepte der Mustererkennung sind robust übertragbar auf unterschiedliche Anwendungsdomänen und lieferten im Rahmen zahlreicher Veröffentlichungen State-of-the-Art Ergebnisse.

[17] `https://birdnet.cornell.edu/`

Das erarbeitete Konzept, DSPs als weit fassbaren Begriff zu behandeln, hat sich als sinnvoll erwiesen und zeigt, dass viele Algorithmen bereits auf kleinen/sparsamen ARM- oder Mobilarchitekturen –und auf GPU ohnehin –echtzeitfähig umsetzbar sind.

2.4.5 Audiobasierte Lokalisation durch Laufzeitunterschiede

In Abstimmung mit der Stifterfirma Intenta GmbH halten wir Lokalisation per Signalstärke für weniger relevant als Lokalisation per Laufzeitunterschied. Daher legten wir die Priorität auf Laufzeitunterschied-basierte Verfahren.

Die geringere Relevanz der Lokalisationsverfahren per Signalstärke ist wiefolgt zu begründen: Die Intenta S2000 Kameras verfügen in der aktuellen Iteration nicht über Mikrofone. Anders als ursprünglich im Projektantrag angedacht, kann daher keine Koppelung der Kameras und Kommunikation bezüglich der Stärke empfangener Audiosignale erfolgen. Infolgedessen kann keine akustische Peilung auf diesem Wege vorgenommen werden. In diesem Zusammenhang wurden die in Kapitel 2.2.3 beschriebenen Testbeds von getrennten und gemischten Video- und Audiodaten mit planaren Mikrofonarrays aufgezeichnet (Abbildung 52). Diese Testbeds sollten für erste Audio-Lokalisationsversuche vorrangig verwendet werden, was in Verbindung mit Signalstärke-basierten Verfahren jedoch wenig praktikabel ist.

Dies liegt darin begründet, dass das Messprinzip auf der Annahme basiert, dass die Art der Schallquelle bekannt ist, diese immer einen konstanten Schallpegel emittiert und dieser Schallpegel als Referenzwert a priori bekannt ist. In den vom Stifter Intenta fokussierten Anwendungskontexten –*Ambient Assisted Living* und öffentliche Verkehrsmittel –findet sich allerdings häufig z.B. menschliche Sprache, welche nicht konstant gleich laut ist und damit die vorgenannten Annahmen nicht erfüllt.

In der Literatur finden sich vereinzelt aktuelle Veröffentlichungen zu Signalstärke-basierten Verfahren [90]. Häufig werden dabei so genannte verteilte Sensornetzwerke, also Netzwerke aus Mikrofonsensoren, verwendet. Wir versprechen uns von Laufzeitunterschied-basierten Verfahren dennoch bessere Ergebnisse.

Wie eingangs erwähnt, haben wir für die initialen Experimente der Audio-Lokalisation das Testbed „16 Statische Lautsprecher-Aufnahmen mit 56 Mikrofonen“ aus Kapitel 2.2.3 verwendet. Dabei handelt es sich, wie in Abbildung 52 zu sehen, zunächst um statische Schallquellen. Zur Ermittlung des Laufzeitunterschieds der Schallquellen bezüglich der Mikrofonpaare haben wir einen Ansatz basierend auf der (Generalized) Cross Correlation gewählt. Im Rahmen der Auswertung der Ergebnisse haben wir verschiedene Hypothesen aufgestellt und untersucht. In diesem Zusammenhang haben wir verifiziert, dass das Aufnahmesystem ordnungsgemäß funktioniert, die Synchronisation fehlerfrei läuft und das Testbed

Abbildung 52: Reales und virtuelles 3D-Schema des verwendeten Testbeds.

die physikalische Realität abbildet. Wie Abbildung 53 zeigt, liefert das Verfahren für das verwendete Testbed eine Richtungsgenauigkeit zwischen ± 5 ° und ± 10 °, je nach Mikrofon-Array. In einem zweitem Schritt könnte ein weiteres Verfahren angewendet werden, um aus den in Abbildung 54 gezeigten geschätzten Raumrichtungen, eine konkrete 3D-Koordinate zu berechnen.

Da jedoch bereits eigene Vorarbeiten für ein Laufzeitunterschied-basiertes Verfahren mithilfe von linearen Gleichungssystemen existierten [129], konzentrierten wir uns auf dieses Verfahren, welches den Vorteil bietet, dass die 3D-Koordinate der zu lokalisierenden Schallquelle direkt in einem einzigen Schritt berechnet wird.

Um das oben genannte Gleichungssystem-Verfahren zu evaluieren, haben wir ein neues Testbed erstellt. Details sind in Kapitel 2.2.3 unter „Saugroboter beim Fahren durch das Labor“ beschrieben. Unter der Maßgabe des von der Stifterfirma Intenta fokussierten *Ambient Assisted Living*-Anwendungsfalls, haben wir einen Saugroboter im Labor umherfahren lassen und diesen lokalisiert. Hierfür änderten wir die Mikrofongeometrie, sodass sich 8 Mikrofone in den Raumecken des Laborkäfigs befanden. Abbildung 55b zeigt diesen Aufbau. Die Erstellung der Groundtruth-3D-Koordinaten des umherfahrenden Saugroboters wurde durch zeitgleiche Aufnahme von Audio- und Videodaten bewerkstelligt. Dieser Prozess ist in Abbildung 55a dargestellt. Es konnte gezeigt werden, dass die Verfolgung des Saugroboters im Raum allein anhand seines Motorengeräuschs möglich ist –aber auch bei gleichzeitiger Wiedergabe von Musik oder Sprache über einen auf dem Saugroboter fixierten Bluetooth-Lautsprecher. Eine visuelle Darstellung des Laborszenarios und der Lokalisationsergebnisse zeigt Abbildung 56. Die Ergebnisse wurden bei der Matlab-Expo im Rahmen eines Vortrags präsentiert [72].

Im Rahmen dieser Arbeitspakete konnte ebenfalls eine Masterarbeit betreut und erfolgreich abgeschlossen werden [28].

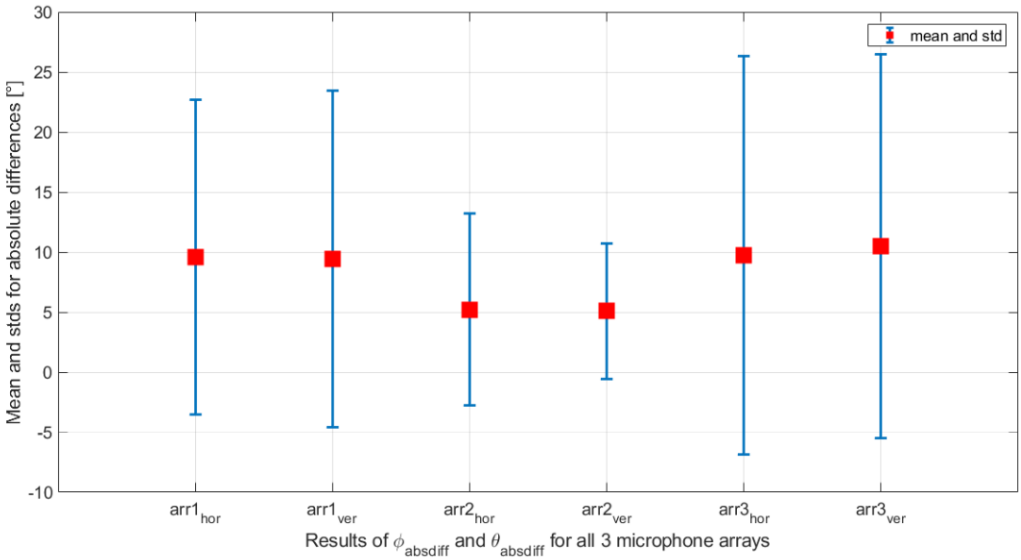


Abbildung 53: Ergebnisse der Evaluation der Audiolokalisation durch Laufzeitunterschied von 3 verschiedenen, flächigen Mikrofon-Arrays. Gezeigt sind Mittelwerte und Standardabweichungen der bestimmten von den echten horizontalen und vertikalen Winkeln zwischen Mikrofonen je eines Arrays und 16 verschieden aufgestellten Schallquellen, wie in Abbildung 52 visualisiert.

2.4.6 Bilanz nach Erreichen des letzten Meilensteins in AB 3

Die Laufzeit-basierte Lokalisierung einzelner Objekte konnte für flächige- wie räumliche Mikrofon-Geometrien gezeigt werden, jedoch stellt *Beamforming* hier die *State-of-the-Art*-Technologie dar. Letztere konnte im Projektkontext aber nicht mehr maßgeblich implementiert und evaluiert werden. Die audiobasierte Lokalisierung und Klassifikation konnte an vielen Beispielen eindrücklich gezeigt werden, vor allem bei der Klassifikation von Vögeln, wo unter 1500 Klassen trotzdem Erkennungsraten von ca. 70 % möglich sind und bei weniger Klassen deutlich höhere Raten [68]. Der gewählte Ansatz über die in Theano und Lasagne entwickelte CNN-Architektur erwies sich als übertragbar auf andere, industrierelevantere Audioklassen wie im *Ambient Assisted Living*, wo 85 % - 90 % *Mean Average Precision* im Bereich von 50 - 100 Klassen möglich sind [108]. Außerdem konnte gezeigt werden, dass sich die parameterreichen CNNs auch reduzieren lassen und damit auf mobiler- und low-budget Hardware, wie einem Raspberry Pi Mini-PC, echtzeitfähig Audiostreams verarbeiten können –unter wenigen % Einbußen in den Erkennungsraten. Zudem zählt die Publikation von Stefan Kahl [68] zu den meistgelesenen und meistzitierten des Projekts, trotz der Tatsache, dass diese erst vor etwa 2 Jahren erschien. Neben dem TRECVID-Wettbewerb hat dies sicher maßgeblich zur internationalen Sichtbarkeit beigetragen.

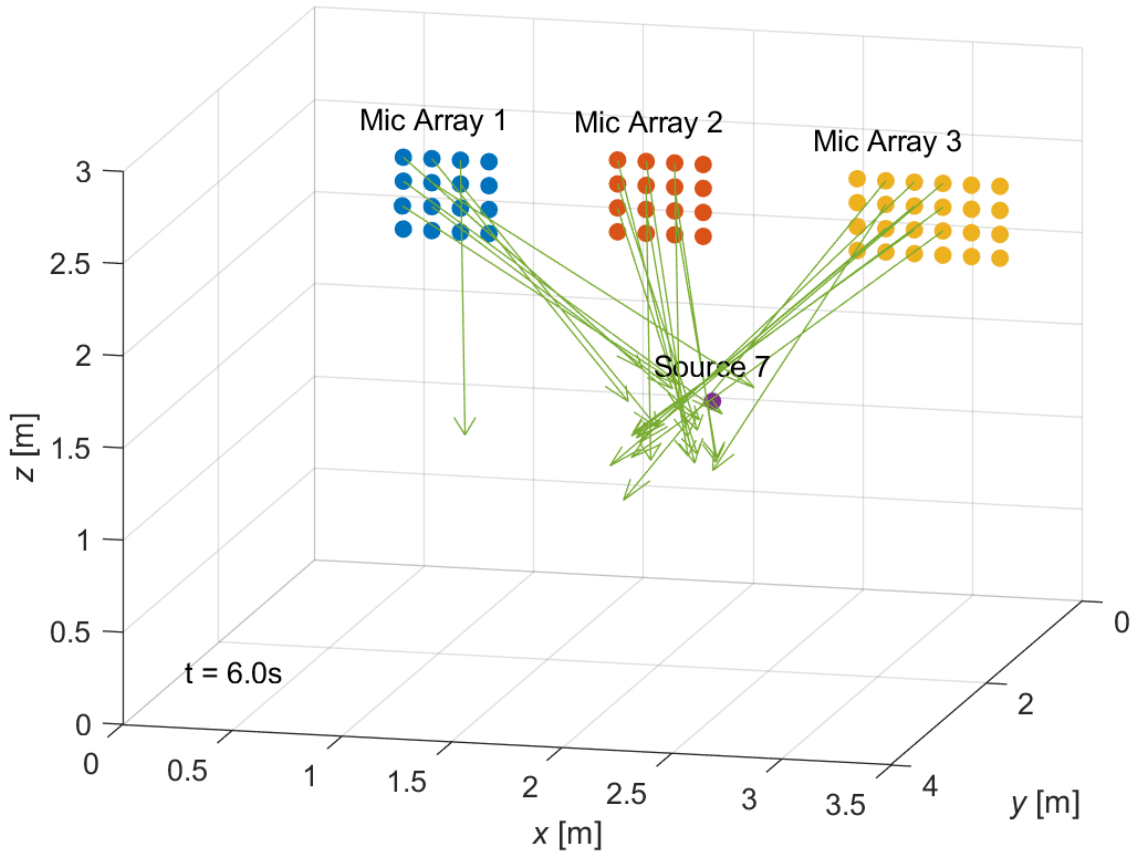


Abbildung 54: 3D-Visualisierung der relativen räumlichen Anordnung von 3 Mikrofon-Arrays und einer lokalisierten Beispiel-Schallquelle (Lautsprecher) zum Zeitpunkt t = 6 s der Aufnahme mit 1 s Fensterbreite. Jeder Pfeil visualisiert den jeweils horizontal- und vertikal berechneten Winkel eines Mikrofonpaares. Es zeigt sich eine deutliche Tendenz zur Schallquelle hin, aber auch die Streuungen, die sich in den Standardabweichungen von Abbildung 53 zeitlich und räumlich gemittelt ergeben.

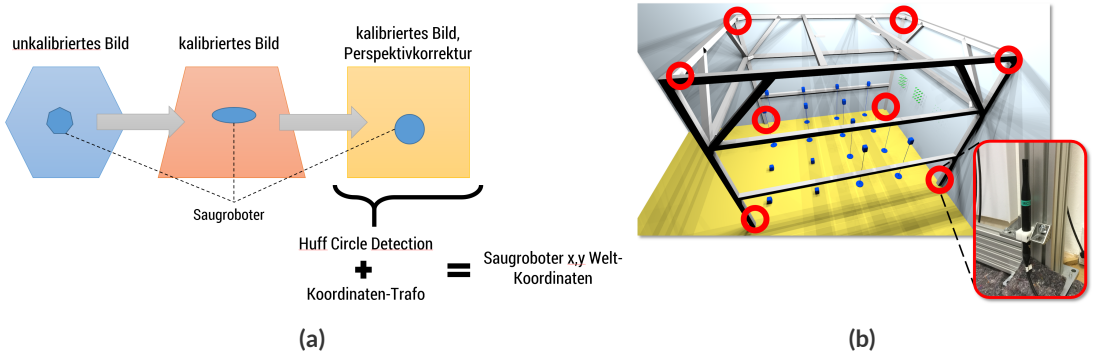


Abbildung 55: a) *Ground Truth*-Erstellung für Saugroboter Testbed mittels Kamera-Sensoren. **b)** Verteilte Mikrofongeometrie mit 8 Mikrofonen in den Raumecken des Laborkäfigs.

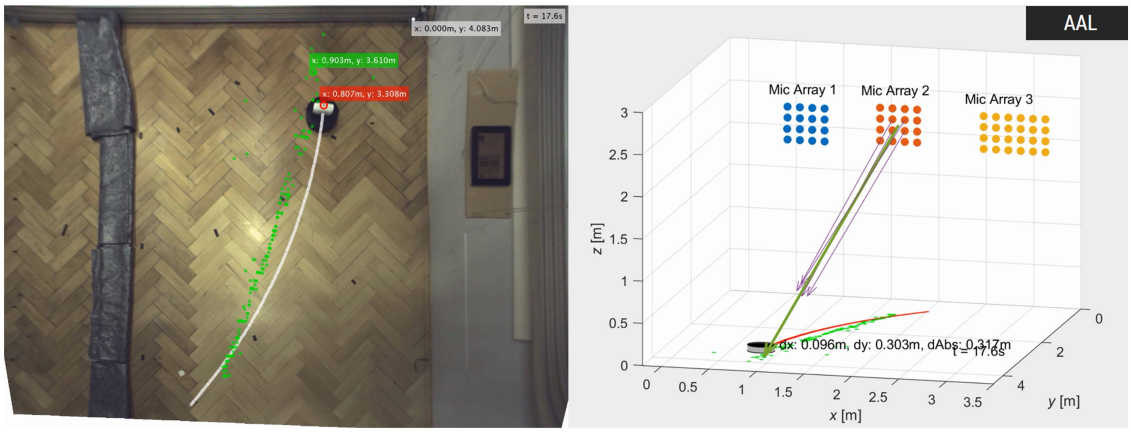


Abbildung 56: Evaluation der Laufzeitunterschied-basierten Lokalisation einer beweglichen Schallquelle. Links: Bluetooth-Lautsprecher auf einem Saugroboter, aufgenommen bei Wiedergabe einer Sprachaufnahme. Weiße Linie - Videobasierte *Ground Truth*-Trajektorie. Grüne Punkte - Bestimmte 3D-Koordinaten aus Audio-Lokalisation. Rechts: Schematische Darstellung des Laborszenarios mit 3 Mikrofon-Arrays und der aus vier Mikrofonpaaren (lila Pfeile) gemittelten Richtung (grüner Pfeil). Die Werte dx, dy und dAbs zeigen die absoluten Fehler in x- und y-Richtung sowie des Abstandes in Metern.

2.5 AB 4: Lokalisierung in Verarbeitungsprozessen

In der Fabrikproduktion werden Werkstücke zunehmend automatisch bearbeitet. Ein Beispiel hierfür stellt die Lasertechnik dar, mit deren Hilfe Werkstücke geschweißt, gefräst und geschnitten werden. Hierfür ist eine sehr genaue visuelle Analyse des Prozesses notwendig, d.h. die aktuelle Position des Werkstückes im Vergleich zum Lasergerät muss sehr genau ermittelt werden, um auch auf kleinstem Raum präzise arbeiten zu können. Dazu muss die Bildanalyse in den Verarbeitungsprozess integriert werden, und gleichzeitig optional ein externes Eingreifen und Steuern ermöglichen. Die Ansätze, welche in dem Vorgängerprojekt sachsMedia entwickelt wurden, sind vor allem für Informationssysteme und den Bereich Entertainment gedacht. Für den Einsatz in der Fabrikation müssen neue Mechanismen erstellt werden.

2.5.1 Aufgabenstellung und Zielsetzung

- Testbeds: (1) Schweißzonenvideos und (2) Halbleiter-Wafer:
  Die konkreten Anwendungsfelder wurden mit dem Stifter 3D-Micromac AG abgestimmt und die dazugehörigen Daten durch diesen oder dessen Partner zur Verfügung gestellt und werden im ersten der folgenden Unterkapitel vorgestellt.

- Analyse laserverarbeitender Prozesse durch Bild und Video:
  - Die im Antrag avisierte Analyse von Werkstück-Position und Schnittvorgangängen zur Prozessüberwachung in laserverarbeitenden Prozessen wurde mit dem Stifter 3D Micromac AG konkretisiert. Dabei wurden die Anforderungen auf die beiden zuvor genannten Testbeds übertragen. (1) Beim Schweißprozess sollte weniger das Werkstück, eine metallische Oberfläche, als die vom Laser verursachte Schmelzzone auf dem Werkstück an sich lokalisiert werden. Dies umfasst vor allem eine zeitlich-räumlich Beschreibung, um anschließend mit den Prozessparametern korreliert zu werden. Eine zeitlich-räumlich Beschreibung impliziert dabei die Aufnahme und Auswertung via eines Videos. (2) Ein weiterer lasergeführter Prozess ist die sogenannte *Thermal Laser Separation* (TLS), ein Prozessschritt bei dem das „Werkstück“, ein Halbleiter-Wafer, in einzelne Chips geschnitten wird. Vereinbartes Ziel war die Qualitätskontrolle dieses Prozesses nach dem Schnittprozess zu evaluieren. Dazu mussten Chips und Schnitte (genannt Straßen) lokalisiert und deren Qualität automatisiert beurteilt werden. TLS ist ein von 3D-Micromac AG zusammen mit Jenoptik und Frauenhofer Institut IISB entwickeltes Verfahren, auf dem die 3D-Micromac AG weltweiter Marktführer ist [18], [81, 130].

[18] https://www.ihk-nuernberg.de/de/Geschaeftsbereiche/Innovation-Umwelt/IuK-E-Business/Mikroelektronik/Innovationspreis/klare-kante

- Berücksichtigung prozessbedingter Bildqualitätseinschränkungen:
  Die zu entwickelnden Algorithmen sollen mit prozessbedingten Einschränkungen der Bildqualität umgehen können. Es ist angestrebt die Analysen zu evaluieren um darauf derartige Einschränkungen eingehen zu können. (1) Durch die hohe Bildwiederholrate (engl., **Framerate**) ist das Signal-Rausch-Verhältnis im Prozess limitiert, aber schwer vermeidbar. Entsprechend muss mussten modellbasierte und zeitliche Mittelungsstrategien entwickelt werden. (2) Einige Fehler durch den TLS-Prozess sind sehr klein und darüber hinaus sind Fehler relativ selten und haben doch diverse Erscheinungsformen. Diese Anforderungen sind entsprechend zu berücksichtigen und wurden im Projekt sehr genau evaluiert. Die Bildqualität schwankt stark zwischen den Bildern, da unterschiedliche Strukturen und Materialien, teils aber auch unterschiedliche Aufnahmesysteme verwendet wurden. Daher wurde versucht ein möglichst generischer algorithmischer Ansatz zu gehen, der gut verallgemeinert auf andere Proben.

- Realisierung von Überwachung und Steuerung:
  In Absprache mit den Stiftern wurde beschlossen, dass der Fokus auf Überwachung von Prozessen und dem TLS liegt und deren Evaluation im Vordergrund steht. Eine Kopplung mit der Prozesssteuerung war nach Bearbeitung des Testbendes 1 nicht vorgesehen.

- Echtzeitfähigkeit:
  Mit Bezug auf den vorherigen Punkt stand die Güte der Algorithmen durch robuste Evaluationen im Vordergrund. Trotzdem wurde die Verarbeitungsgeschwindigkeit stets mit untersucht.

- Schnittstellendefinition und grafische Nutzeroberflächen:
  Initial wurde im Projekt eine generische grafische Nutzeroberfläche für Laserverarbeitende Maschinen entwickelt. Aufgrund der angepassten Prioritäten wurde dieser Ansatz über den Mockup hinaus aber nicht weiterentwickelt. Für Thema (1) wurde eine eine Nutzeroberfläche entwickelt, welche die Prozessschritte auch für neue Datensätze wiederholbar macht und kein Programmierwissen voraussetzt. (2) Im Thema 2 lag der Schwerpunkt eher darin zu mehr annotierten Daten zu kommen. Daher wurden die im Projekte entwickelten Annotationstools angepasst um gemeinsam mit den Prozessingenieure *Ground Truth* erzeugen zu können.

2.5.2 Testbeds: Schweißzonenvideos und Halbleiter-Wafer

(1) Schweißzonenvideos: Vom Stifter 3D-Micromac AG wurden ca. 100 monochrome Videos im *.cine-Format zur Verfügung gestellt, die Laserschweißprozesse zeigen, aufgenom-

men mit einer Hochgeschwindigkeitskamera. Jedes dieser Videos hat eine einzigartige Prozessparameterierung, die als Tabelle zur Verfügung gestellt wurde. Zu den variierten Parametern gehörten, die Leistung des Lasers, die Bewegungsgeschwindikeit der Düse, Gasstromparameter und Kameraperspektive. Die Daten wurden für die Verarbeitung in .mp4 konvertiert. Die Erstellung der Aufnahmen erfolgte an der Professur Schweißtechnik der Technischen Universität Chemnitz im Rahmen der Doktorarbeit von Björn John, der die Forschung im Projekt LocalizeIT begleitete [58]. (2) Halbleiter-Wafer: Vom Stifter 3D-Micromac AG wurden 10 hochaufgelöste Bilder verschiedener Halbleiter-Wafer im jpg-Format zur Verfügung gestellt. Zu sehen sind darauf einzelne Chips und dazwischen Schnittgrenzen, sog. Straßen, die vor dem Vereinzeln der Chips geschnitten werden. Diese Bilder werden direkt nach dem TLS Dicing mit mikroskopischer Auflösung aufgenommen und zusammengesetzt. Dabei entsteht ein hoch aufgelöstes Bild des Wafers mit lasergeschnittenen Sägestraßen. Ein solches Wafer-Bild ist links in Abbildung 62 zu sehen. Neben dem Gesamtbild wurde ein weiterer Datensatz der gleichen Wafer bereitgestellt, jedoch aufgeteilt in einzelne Chips so wie sie das Aufnahmesystem vor dem *Stitching* generiert. Zu sehen sind darin Nahaufnahmen von einzelnen Schnittregionen. Zusätzlich wurden für 5 Wafer Annotationen von Prozessingenieuren zur Verfügung gestellt, welche die Qualität der einzelnen Chips und Straßen bewerten. Neben mehrheitlich „guten“ Chips wurden auch lokale Fehler bzw. Auffälligkeiten annotiert und als Tabelle zur Verfügung gestellt. Die Gesamtgröße des Datensatzes beträgt 3GB.

2.5.3 Analyse laserverarbeitender Prozesse durch Bild und Video - Schweißzonenanalyse

Das Videomaterial, das in Kapitel 2.5.2 bereits grob vorgestellt wurde, stammt aus dem Bereich der Lasermikrobearbeitung gewählt, welches das Schweißen metallischer Werkstoffe mittels Laserstrahl darstellt. In diesem Bereich stellen sich besondere Herausforderungen durch die Werkstoffdicke, um einen erfolgreichen Fügeprozess sicherzustellen. Durch die Manipulation von verschiedenen Faktoren des Laserschweißsystems kann die Qualität erhöht werden. Zum Beispiel kann durch Betrachtung der Schmelzzone mittels Hochgeschwindigkeitesvideoaufnahmen eine Aussage über das Prozessverhaltens und somit die Qualität der Fügeverbindung getroffen werden. Im Rahmen von AB4 wurde mit Methoden der Bildverarbeitung die Auswirkungen unterschiedlicher Bedingungen auf die Schmelzzone und damit der Fügeverbindung analysiert und charakterisiert. Dazu wurde ein Werkzeug realisiert, das das einlesen der Videodaten ermöglicht und in den Speicher des Computers einliest. Die im Speicher vorhandenen Daten werden nun auf Konsistenz und Anforderungen geprüft. Umgesetzte Anforderungen sind zum Beispiel die korrekte Bildauflösung, *Framerate* und Dateigröße. Mit Hilfe von Algorithmen findet die Analyse und Charakteri-

sierung auf den auf Konsistenz geprüften Daten statt und anschließend eine Visualisierung durch eine grafische Programmoberfläche (GUI).

Der für die Schmelzzonen Analyse implementierte Algorithmus ist in Bild 58 schematisch dargestellt und wurde in das entwickelte GUI-basierte Funktionsmuster eingebettet. Bild 57 zeigt am Beispiel eines Bildes aus der Videodatei, in welcher Reihenfolge der Algorithmus für die Bildverarbeitung abläuft und welche Ergebnisse dabei schrittweise entstehen. Die geometrische Approximation der Schmelzzone, schematisch dargestellt in Bild 58 nutzt zwei gegebene, konsekutive Einzelbilder als Grundlage, Weichzeichneroperatoren verschiedener Größen und Kantendetektoren. Dadurch wird Rauschen unterdrückt und es können signifikanten Kanten detektiert werden. Durch Fusion dieser beiden Element mengen bildet sich ein Kantendifferenzbild, das durch morphologische Operationen und die Auswahl der stärksten Kanten geglättet und bereinigt wird. Durch einen Konturdetektors werden die inhaltsbeschreibenden Ellipsen definiert, die nach erneuter Filterung das Ergebnis darstellen. Der beschriebene Algorithmus entspricht der umgesetzten Lösung. Es wurden weitere Filter für das Anwendungsbeispiel ausprobiert, die sich nicht bewährten. Weiterhin wurden verschiedene Performanzsteigerungen beim Speicherverbrauch und beim Laufzeitverhalten der Software erreicht. Die Software MATLAB wurde zum *Rapid Prototyping* verwendet, insbesondere um mathematische Operationen zu evaluieren, in der finalen GUI aber neu implementiert in C#.

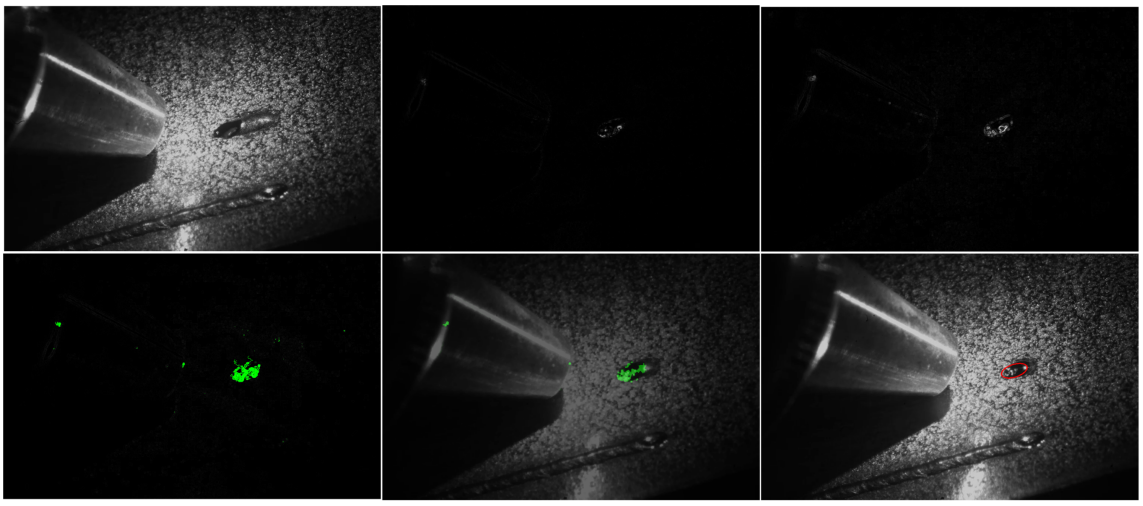


Abbildung 57: Ablauf der Bildverarbeitung am Beispiel

In Abbildung 57 ist in der ersten Reihe links das Ausgangsbild, gefolgt von dem Differenzbild und dem Sobel Kanten Bild. Reihe zwei beginnt auf der linken Seite mit dem Ergebnis nach dem kombinierten Sobel Canny Filter, gefolgt von dem Ergebnis nach der morpholgischen Operation (zusammenhängende Objekte) und schließt mit dem Ergebnisbild ab. Als Pro-

grammiersprache für das Werkzeug wurde C# verwendet, auch um existierende Funktionen aus der großen Community und das einfache Erweiterbarkeit bzw. Erlernbarkeit durch andere Teammitglieder sicher zu stellen.

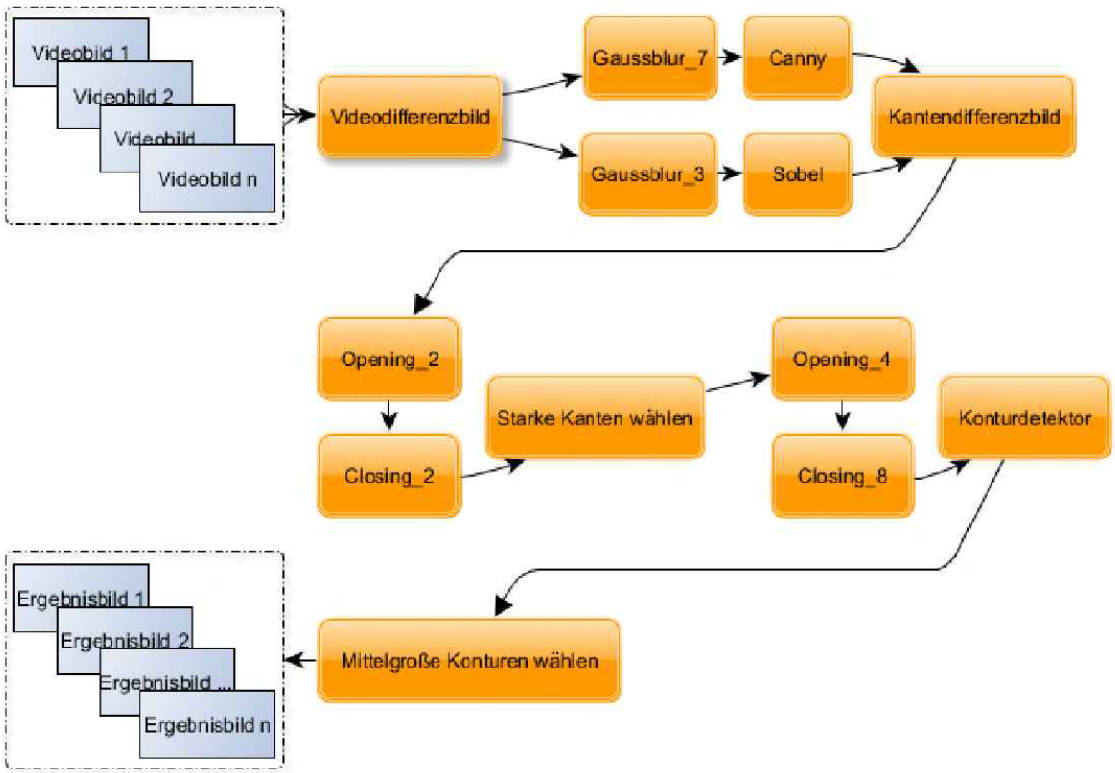


Abbildung 58: Algorithmus für Schmelzonenanalyse [102]

In Bild 59 ist ein Screenshot des Werkzeuges zu sehen. Dabei ist auf der linken Seite ein Bild aus dem Video zu sehen, in dem das Ergebnis, die rote Markierung, als Schmelzzone detektiert wurde. Die rechte Seite zeigt aktuelle Abarbeitungsschritte und den Fortschritt an. An der oberen Seite der Software sind drei *Tabs* zu sehen. Der *Tab* „Video Processing“ zeigt, wie im Bild die Abarbeitung an, wogegen der *Tab* „Curve Fitting“ die grafische Darstellung der Evaluation darstellt (wie in Bild 60) und der *Tabs* „Data Visualization“ dem Nutzer ermöglicht Diagramme mit eigener und dynamischer Parametrisierung zu erstellen.

In Bild 60 werden verschieden Diagramme dargestellt, die Ergebnisse aus der Analyse des entwickelten Werkzeuges sind.

In den ersten beiden Projektjahren wurde das Thema videobasierter in situ-Charakterisierung von Schmelzprozessen umfänglich bearbeitet und mit 2 Publikationen auf einer nationalen und einer internationalen Konferenz abgeschlossen [59, 102]. Dabei ist auch ein frei

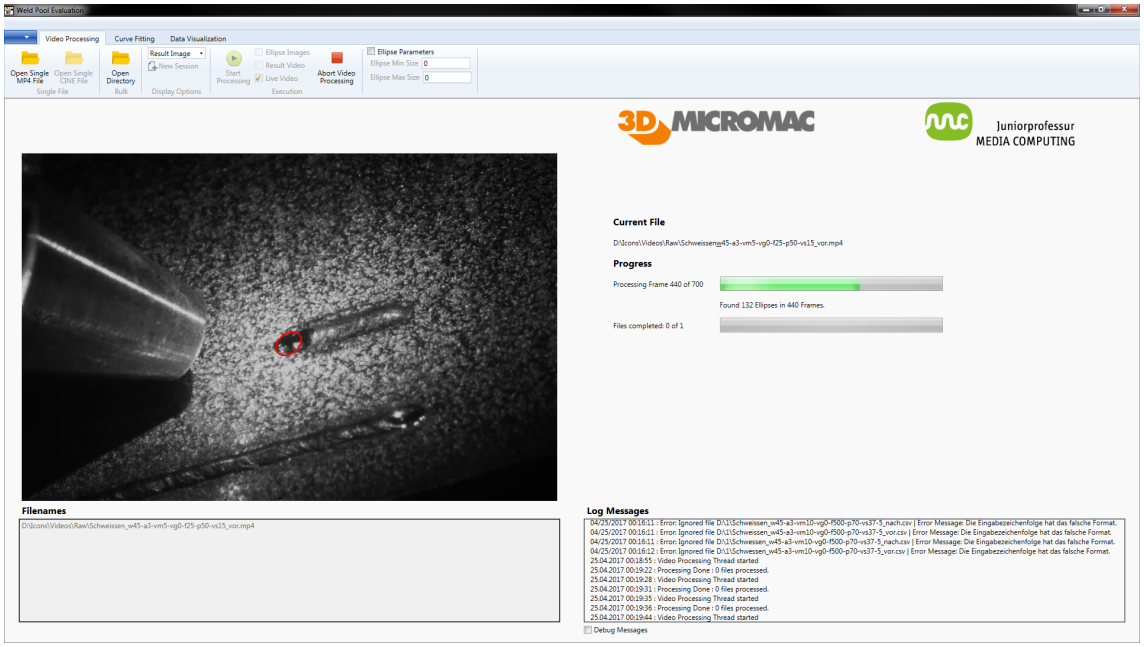


Abbildung 59: Bildschirmfoto des Werkzeuges für Schmelzzonen Analyse

erhältliches Anwenderprogramm entstanden, das der Community zur Verfügung steht [59], zu finden unter dem Name „Weldpool Evaluation Tool“ auf der Projektwebseite[19].

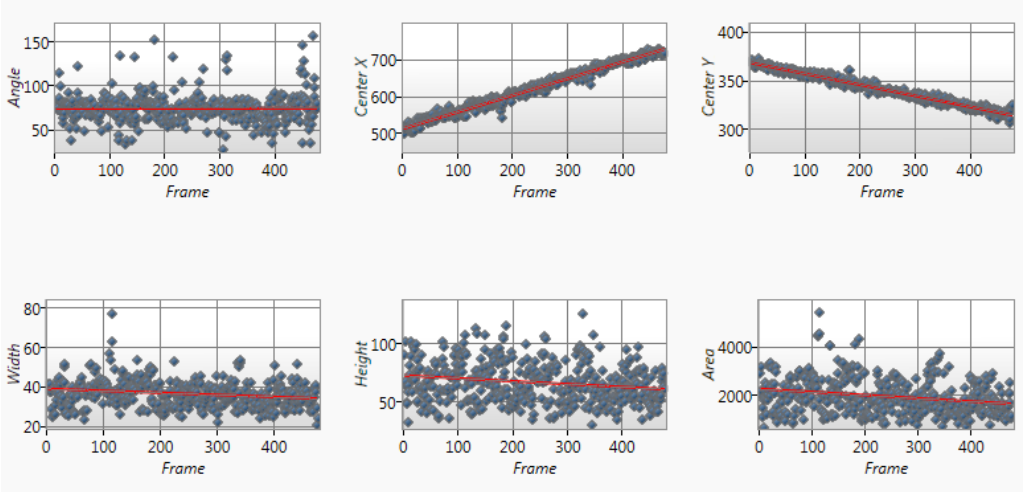


Abbildung 60: Evaluationsergebnisse des Ellipsenparameters

[19] https://localize-it.de/downloads/

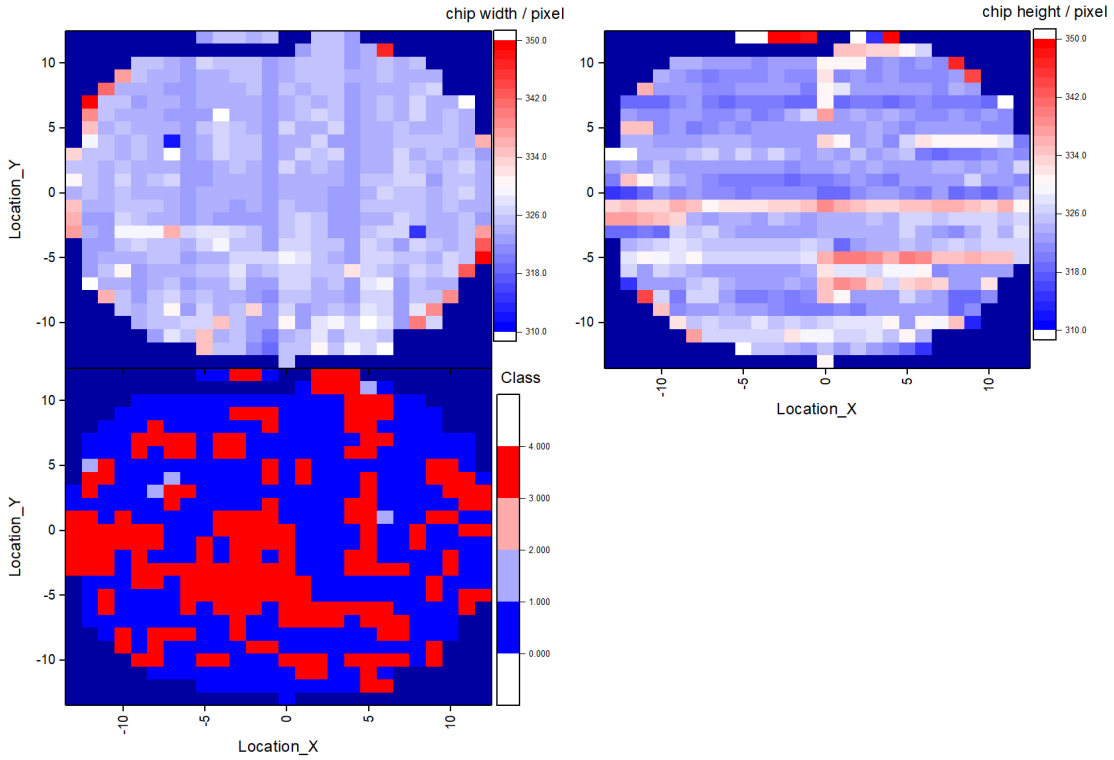


Abbildung 61: Schematische Darstellung der Breite (links, oben) und Höhe (rechts, oben) einzelner Chips nach dem TLS-Dicing. Darunter befindet sich die Klassifikation der Chips mit folgender Zuordnung: 1 (blau) - gut, 2 (hellblau) - interpoliert (gut), 4 (rot) - defekt

2.5.4 Schnittfehler-Analyse auf Halbleiter-Wafern im TLS Schnittverfahren (Thermal Laser Separation)

Eine visuelle Inspektion befasst sich unter dem Einsatz bildverarbeitender Verfahren mit der Detektion und Klassifikation von Herstellungsfehlern in der Halbleiterindustrie, hier genauer im TLS Schnittverfahren. Während eine frühestmögliche Erkennung von Fehlermustern eine Qualitätskontrolle und Automatisierung von Herstellungsketten erlaubt, können Hersteller von einer Erhöhung der Ausbeute und der Reduktion von Herstellungskosten profitieren. In Absprache mit dem Stifter sollte daher im weiteren Projektverlauf das entstehende Produkt des Laser-Verarbeitungsprozesses untersucht werden. Im Falle eines Laserschnittprozesses sind das häufig Wafer aus der Halbleiterindustrie. Diese werden mit dem sogenannte TLS Dicing (Thermal Laser Separation) Verfahren linienweise horizontal und vertikal geschnitten mit dem Ziel die darauf befindlichen Chips zu vereinzeln.

Aus der Analyse der Wafer mit regelbasierten Algorithmen lässt sich feststellen, dass die Verteilung der ermittelten Chip-Breiten und Chip-Höhen räumlich nicht homogen ist, was

beim Sägeprozess auch zu erwarten ist. Eine Heatmap-Darstellung ist in Abbildung 61 dargestellt. Die Breite und Höhe lässt darauf schließen in welchen Straßen die Laser-Steuerung mehr oder weniger Überstand produziert. Idealerweise bleibt die Breite der Chips in vertikalen Chipreihen konstant (gleiche Farbe) und die Höhe in horizontalen Schnittlinien. Aber es wird deutlich, dass nur Teilbereiche diese Erwartung erfüllen. Eine mittige Führung des Lasers durch die Sägestraßen könnte damit optimiert werden. Die Evaluation im dritten Jahr zeigte, dass die Lokalisation einzelner Chips sich als sehr genau erwies, während die Evaluation der Klassifikationsgüte hingegen Schwächen offenbarte, insbesondere bei einem Wechsel zu einem anderen Wafer mit geänderter Struktur und anderen Aufnahmemodalitäten wie Optik, Belichtung und Bildauflösung.

Wafer-Analyse mit modernen Methoden künstlicher Intelligenz

Weil klassische Bildverarbeitungsansätze, wie im vorheriger Abschnitt, in ihren Fähigkeiten oft limitiert sind, verfolgten die weiteren Arbeiten dieses Arbeitsbereiches die Erweiterung bisheriger Verfahren um einen neuartigen, auf *Deep Neural Networks* basierenden hybriden Ansatz. Wir möchten hier insbesondere auch Augenmerk auf einen hybriden Ansatz legen, da die Ansätze des maschinellen Lernens häufig mit einem enormen Rechenaufwand verbunden sind und daran scheitern feinste Strukturen zu erkennen. Im Gegensatz zu klassischen *Deep Neural Networks* sieht der verfolgte hybride Ansatz ein mehrstufiges Verfahren vor, das in hochaufgelöstem Bildmaterial eine Erkennung von feinsten Strukturen in Pixelgröße erlaubt.

Die zur Verarbeitung von Wafern sowie deren Analyse werden im Folgenden im Kontext der Verarbeitung höchst diverser Datenmengen und Anforderungen erläutert. So weisen die Strukturen und Unterstrukturen von Wafern, unterteilt in Chips und Straßen (Abbildung 62), oftmals komplexere Fehlermuster auf, die bereits in Pixelgröße zu identifizieren und zu unterscheiden sind. Bisherige maschinell unterstützte Systeme sahen hierzu eine Kontrolle durch Inspekteure vor, deren Fehleranfälligkeit aufgrund menschlicher Unachtsamkeit, Erschöpfung, möglicher händischer Parametrierungen des zugrundeliegende Systems oder auftretende maschinelle Probleme oftmals einen entscheidenden Faktor darstellte und mit einem enormen händischen Zeitaufwand verbunden war.

Im Folgenden haben wir einmal den Ansatz Richtung einer industriellen Anwendung ausgewertet, und einmal in Richtung der wissenschaftlichen Verbesserung des Forschungsstandes in der Computervision.

Wafer-Analyse mit hybriden Deep Learning Netzen - Industrielles System

Das Ziel war es somit, ein industrielles System zur automatischen Inspektion zu entwickeln und den Schritt der manuellen Inspektion möglichst zu reduzieren. Der durch neuronale Netze biologisch inspirierte Ansatz begründet sich zum einen durch jüngste Entwick-

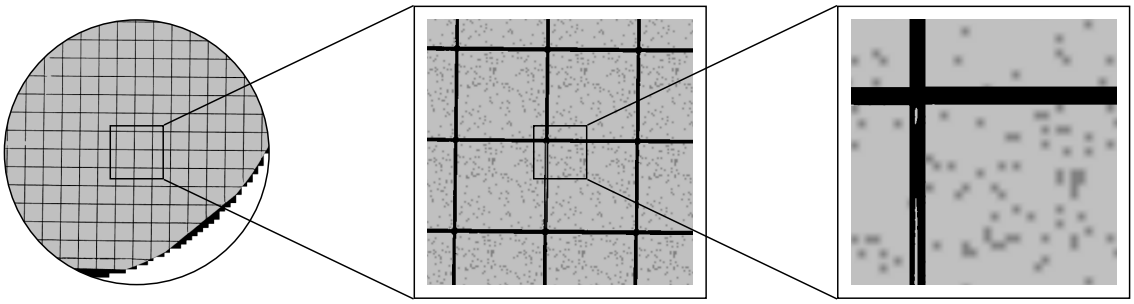


Abbildung 62: Übersicht eines Wafers mit Chip- und Straßen-Ausschnitt (aus urheberrechtlichen Gründen nur umriss-schematisiert dargestellt).

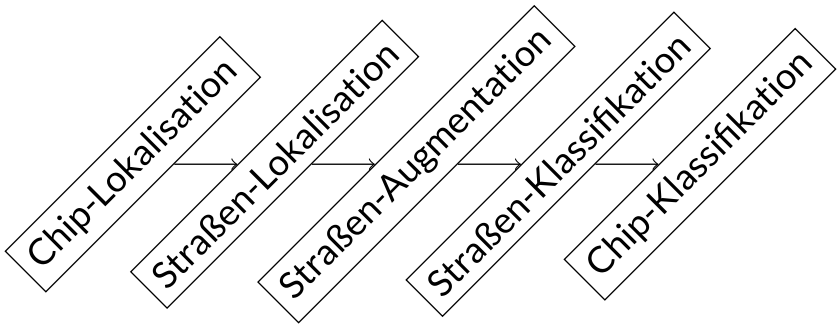


Abbildung 63: Verarbeitungsschritte der Lokalisation, Augmentation und Klassifikation von Chips und Straßen

lungen im Bereich des maschinellen Lernens sowie der Gehirnwissenschaften, aber auch durch wegweisende Forschung der letzten Jahre aus dem Bereich der Halbleiterindustrie. Die aufgenommenen Bilddaten werden klassifiziert, und es stellte sich heraus das mögliche Fehler sehr heterogen sind, wie beispielsweise in Form von kleinen Löchern, Kratzern oder Blasen. Die aufgrund ihrer Diversität bestehenden Wafer- und Fehlerklassen indizieren indessen die Komplexität des Vorhabens, in einer Menge unterschiedlichster Schaltkreise, die durch eine Menge von Strukturen charakterisiert sind, verschiedenartig gute, fehlerhafte oder auffällige Klassen von Mustern bestimmen zu können.

Während sich die Erkennung von Fehlermustern in feinsten Strukturen in Pixelgröße in Abhängigkeit der bestehenden Bildauflösung darstellt, erfolgt zunächst eine Unterscheidung in einem Detailgrad demnach Chips und Straßen separat betrachtet werden. Die folgende schematische Übersicht (Abbildung 63) stellt die hierfür realisierten Schritte der Verarbeitung dar. Dazu werden falls nötig zunächst die Aufnahmen der Wafer entsprechend ihrer Straßen zugeschnitten und eingeordnet. Anschließend erfolgt eine Unterscheidung der separierten Chips in Chips, die sich bereits außerhalb der *Region of Interest* (ROI) befinden, oder vom Rand des Wafers selbst geschnitten werden und somit fehlerhaft sind. Um möglichen geringeren Auftretenshäufigkeiten einzelner Fehlerklassen entgegenwirken zu können, dient eine *Data Augmentation* der Erzeugung neuartig fehlerhafter Straßen. Letztend-

Test run	Mean accuracy
RFC	0.600 ± 0.005
SVC with linear kernel	0.677 ± 0.001
SVC with RBF kernel	0.696 ± 0.000
MLP	0.681 ± 0.020
CNN	$\mathbf{0.757 \pm 0.032}$
SH-CNN	
$1\times$	0.896 ± 0.013
$2\times$	$\mathbf{0.909 \pm 0.011}$
$4\times$	0.880 ± 0.022

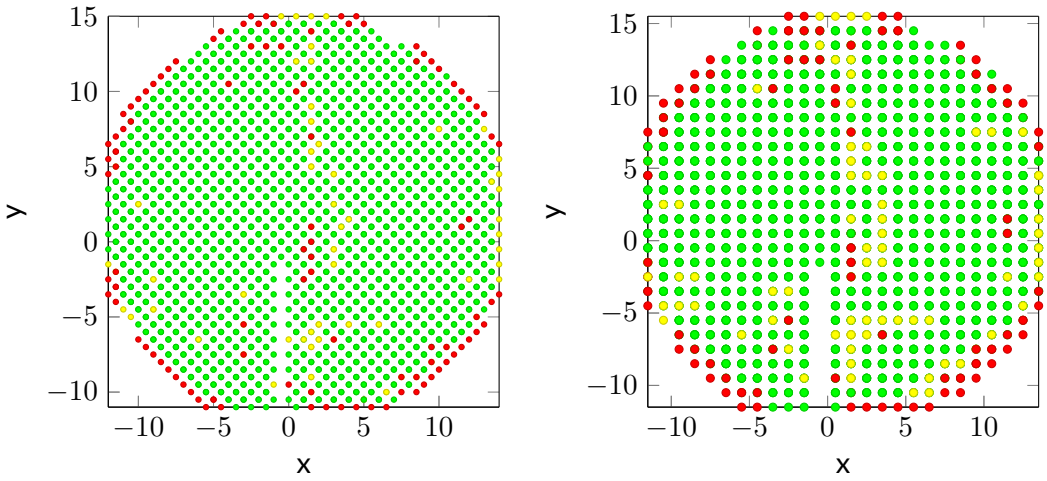
Tabelle 6: Testergebnisse der Straßen-Klassifikation (Datensatz enthält Chips im Wafer-Inneren und auf dem Rand). SH-CNN: eigenes System. Basisverfahren: Random Forest Classifier (RFC), Support Vector Machine Classifier (SVM) mit linear und RBF -Kernel, Multilayer Perceptron (MLP).

lich erfolgt die Klassifikation der Straßen, von denen aus Informationen über die Beschaffenheit der Chips geschlossen wird, und diese somit als fehlerhaft oder nicht klassifiziert werden.

Um die Detektions- und Klassifikationsfähigkeiten von Fehlermustern des realisierten Systems über die Chips und Straßen von Wafern quantifizieren zu können, erfolgte eine erste Evaluation der einzelnen Verarbeitungsschritte. Dies umfasst entsprechend Abbildung 63 die Chip- und Straßen-Lokalisation, gefolgt von der Klassifikation der extrahierten Straßen- und Chip-Fehlerklassen. So weist ein Vergleich klassischer auf CNNs basierender Systeme mit dem realisierten hybriden System bereits eine um ca. 17 % erhöhte Genauigkeit (engl. Accuracy) für die Klassifikation von Chip-Fehlerklassen auf (Tabelle 6). Wie die Ergebnisse der zum Einsatz kommenden Testumgebung zeigen, übertrifft das realisierte hybride System die Ansätze bereits bestehender Verfahren, wobei eine Unterscheidung in Abhängigkeit des Detailgrades die Bestimmung und Behebung von Fehlermustern bereits in frühen Stadien des Herstellungsprozesses ermöglicht.

Die anschließende Visualisierung der Straßen- und Chip-Fehlerklassen wurde dann für die bestimmten Ergebnisse anhand des Originals-Wafers realisiert, dessen Straßen und Chips klassenabhängig eingefärbt und dem Anwender zur weiteren Inspektion dargestellt werden können (Abbildung 64). Dies ist insbesondere nötig für ein Produktionssystem.

Der vorgestellte Beitrag wurde zunächst im Rahmen des wissenschaftlichen Projektworkshops bei den Chemnitzer Linux-Tagen 2019 publiziert [111], und dann auf der interna-



(a) Straßen-Klassifikation (b) Chip-Klassifikation

Abbildung 64: Testergebnisse der Straßen- und Chip-Klassifikation visualisiert für fehlerfreie (●), auffällige (●) und fehlerhafte (●) Straßen und Chips

tionalen Maschinenbau-Konferenz „IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)“ publiziert [110], welche die Sichtbarkeit in der Maschinenbau-Industrie Community ermöglicht.

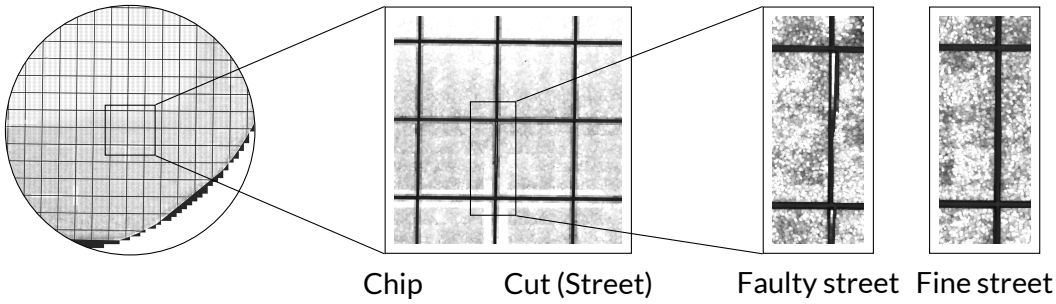


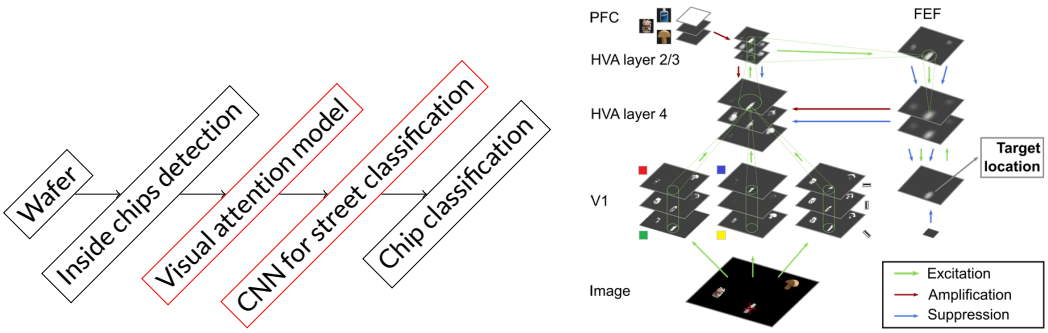
Abbildung 65: Überblick über einen Wafer (Links), Chips (Mitte) und Chipfehler (Rechts) verursacht von dem Schnittprozess. Die Schnitte durch den Wafer werden Straßen (Streets) genannt.

Wissenschaftliche Auswertung des Themas - Einreichung zur International Conference on Computer Vision (ICCV) und Conference on Computer Vision and Pattern Recognition (CVPR)

Der hybride Ansatz, zentriert um den Detailgrad, ist auch für die Computervision Community sehr interessant, da er eine Aufgabe löst die Deep Neural Networks nicht können. Um die wissenschaftliche Auswertung des Projektes zu erörtern, wurde das Vorhaben zusätzlich detailliert evaluiert und auf internationalen Konferenzen eingereicht. Dies resultierte in einer Einreichung auf der International Conference on Computer Vision (ICCV) 2019. Die ICCV ist eine der beiden Top-Level Konferenzen im Bereich CV und wissenschaftlich auf dem höchsten Level gerankt (A nach ERA, A1 nach Qualis). Wir versprachen uns von einer Teilnahme eine erstklassige Publikation und international weithin sichtbare Verbreitung der Projektergebnisse. Folgenden ist das Vorhaben kurz geschildert:

Titel: **„Improving wafer fault detection by combining a biologically plausible model of visual attention with deep learning**“

Es ist ein langfristiges Ziel in den Computerwissenschaften biologische Verarbeitungsprinzipien und die Erkennungskraft des menschlichen Sehsystems auf maschinelle Systeme zu transferieren. In diesem Bereich wird ein Prinzip, die menschliche visuelle Aufmerksamkeit, untersucht anhand des Beispiels des Laserschneidens. Der eingereichte Beitrag schlägt einen neuen Ansatz vor wie mittels Einsatz biologisch-motivierter Verarbeitungsprinzipien visueller Aufmerksamkeit die Erkennung von Fehlern in der Halbleiterindustrie verbessert werden kann. Es wird in der Arbeit die Thematik des Schneidens von Wafern in einzelne Chips (*Semiconductor Wafer Dicing*) untersucht, in welchem beim Schneidprozess auftretende Fehler automatisch und visuell erkannt werden müssen (Abbildung 65). Die automatische Detektion solcher Fehler ist ein sehr wichtiges Thema in der Halbleiterindustrie, da es die Zeit verringert um einen Chip zu inspizieren, den Arbeitsaufwand reduziert und somit die Gewinne für die Hersteller erhöht. Bisherige Arbeiten in der Domain benutzten oft



(a) Komplett-System (b) Visuelles Aufmerksamkeitsmodell

Abbildung 66: a) Komplett-System, mit seinen Kernkomponenten: visuellen Aufmerksamkeitsmodell und *Convolutional Neuronal Network* (CNN).
b) Das biologisch-plausible, visuelle Aufmerksamkeitsmodell. Es ist hier exemplarisch in der Aufgabe vorgestellt ein Objekt, die Flasche, zu suchen (Abbildung aus [26]). In der vorliegenden Anwendung würde es eine Chip-Kante (Straße) suchen. Gehirnareale: V1: primärer visueller Cortex, ein frühes visuelles Areal. HVA: höheres visuelles Areal, vergleichbar mit den Gehirn-Cortices V4 oder IT. HVA ist aufgeteilt um neurophysiologische Daten zu replizieren in separate, kortikale Schichten (Schicht 4 und Schicht 2/3). PFC: Prefrontaler Cortex, enkodiert Objektkategorien. FEF: Frontales Augenfeld, involviert in Verarbeitung von Ortsinformationen.

klassische CV-Ansätze, und nur einige Arbeiten verwenden neuere Ansätze wie Deep Learning [78], u.a. [79, 80, 92, 36]. Ein Problem in der Domain ist das die Fehler winzig sind und innerhalb eines viel größeren Bildmaterials wie des Bildes eines Chips oder ganzen Wafer detektiert werden müssen (Abbildung 65). Daher, es geht um das Problem kleine Strukturen in einem großen Datenmaterial zu detektieren und erkennen.

Ein interessantes Prinzip für dieses Problem ist visuelle Aufmerksamkeit, ein smartes menschliches Prinzip das Verarbeitungsressourcen auf einen aufgaben-relevanten Aspekt der Szene konzentriert [53]. Das ist quasi der theoretische Hintergrund für das in Abschnitt 2.5.4 geschilderte System. In anderen Domänen existieren bereits einige wenige Ansätze um visuelle Aufmerksamkeit mit modernen *Machine Learning* Ansätzen wie Deep Learning zu kombinieren. Allerdings erscheinen die Kombinationen recht willkürlich, sie sind weit entfernt von neurowissenschaftlichen Erkenntnissen, und es erscheint somit nicht wirklich klar wie wirklich kombiniert werden sollte. Daher schlagen wir hier vor ein biologisches-plausibles Modell von visueller Aufmerksamkeit aus den Neurowissenschaften zu verwenden. Das Modell ruht auf einem breiten Fundus neurowissenschaftlicher Daten, es kann so-

Approach	Accuracy [%]	Fault detect. accuracy [%]
Baseline		
KNN	74.97 ± 0.00	65.00
SVM	74.19 ± 0.00	63.00
MLP	68.96 ± 5.06	64.80
ResNet50*[54]	78.76 ± 2.76	60.60
CNN [92]	75.24 ± 1.88	57.20
CNN [36]	78.89 ± 3.03	62.20
Our CNN	80.83 ± 2.38	67.40
Attention-based		
Attention + KNN	72.33 ± 0.00	47.00
Attention + SVM	75.06 ± 0.00	58.00
Attention + MLP	69.26 ± 1.20	46.40
Attention + ResNet50*	81.35 ± 0.50	63.60
Attention + CNN [92]	88.66 ± 0.88	80.40
Attention + CNN [36]	87.63 ± 1.20	76.80
Attention + Our CNN	$\mathbf{91.91 \pm 0.57}$	**87.80**

Tabelle 7: Erkennungsgenauigkeiten für (i) Baseline-Lösungen (d.h. basierend auf Chip Daten), und (ii) Aufmerksamkeits-basierte Ansätze (d.h. basierend auf Straßen Daten). CNN [92], [36] stellen den State-of-the-Art in der Wafer-Domain dar. *Transfer-Learning.

wohl multiple neuronale Feuerratenveränderung durch Aufmerksamkeit replizieren [27], als auch menschliche Verhaltensdaten erklären da es auf vorherigen Modellen der Visuelle Suche basiert [53] und auch neue Daten anfittet (OSM, Kap. 5 in [26]). Des Weiteren zeigen erste Arbeiten die Anwendbarkeit des Modells auf Real-Welt Daten [22, 57]. Daraus ergibt sich unser hybrides System (Abbildung 66a), basierend auf visueller Aufmerksamkeit und einem *Convolutional Neuronal Network* (CNN, ein Deep Learning Ansatz). In dem System selektiert das Aufmerksamkeitsmodell (Abbildung 66b) eine interessante ROI für das CNN, welches dann die Daten klassifiziert.

Das System wurde dahingehend evaluiert wie stark der Vorteil von visueller Aufmerksamkeit für ein Deep Netz ist. Davor wurde aber noch untersucht ob das Aufmerksamkeitsmodell überhaupt korrekt arbeitet. Für letzteres wurde einerseits (a) analysiert wie gut das Modell in der vorliegenden Aufgabe menschliches Verhalten zeigt, d.h. wie gut es die Straßen via Augenbewegungen finden kann: Das Modell erreicht eine sehr gute Genauigkeit von $98,5$%. Zusätzlich wurde anderseits (b) evaluiert wie präzise das Modell die Region-

of-Interests (ROIs) extrahiert: Das Modell hatte im Mittel eine Abweichung von ca. $\pm 0,5$ Pixeln, mit einer Standardabweichung von ca. $3,5$ Pixeln. Anschließend wurde untersucht wie sehr Aufmerksamkeit für Deep Netze hilft, dazu wurde die Erkennungsperformanz von einem CNN mit Aufmerksamkeit mit einem CNN ohne Aufmerksamkeit verglichen (Tabelle 7). Das CNN mit Aufmerksamkeit erreicht eine 11% höhere Genauigkeit die Chips korrekt zu klassifizieren als das System ohne Aufmerksamkeit (91.9% statt 80.8%), und viel wichtiger, die Genauigkeit die Chipfehler auch korrekt zu erkennen stieg von 67.4% auf 87.8%. Diese Ergebnisse zeigen das visuelle Aufmerksamkeit stark die Leistung verbessert und das sich der Einsatz biologischer Verarbeitungsprinzipien lohnt. Die Wafer-Domain verwendet überwiegend noch klassische Verfahren der Künstlichen Intelligenz wie SVM, MLPs, etc., und nur sehr wenige Deep Learning Verfahren. Nichtsdestotrotz haben wir unser Verfahren auch gegen die existierenden Deep Neural Netze aus der Domain gebenchmarkt, neben verschiedenen anderen Basisverfahren wie Random Forest Classifier (RFC), Support Vector Machine Classifier (SVM) mit linear und RBF -Kernel, Multilayer Perceptron (MLP) und Standard Convolutional Neural Networks aus der Objekterkennungs-Domain. Die Resultate der erweiterten Evaluation zeigen, dass insbesondere unser hybrider Ansatz, alle anderen Ansätze um Klassen schlägt.

Leider wurde der Beitrag abgelehnt, mit den Hauptargumenten, dass (i) die Auswertungen nicht ausreichend seien und (ii) die eingereichte Publikation für die ICCV nicht breitengültig genug ist, obwohl es ein sehr solider und guter Workshopbeitrag wäre. Bezüglich ersterem haben wir es viel breiter evaluiert, die hier gezeigten Resultate beziehen sich bereits auf die überarbeitete Version. Die originale Version hatte nur 2 der 10 Resultate. Zweitens haben wir es viel allgemeingültiger geschrieben, in dem wir den Teil bzgl. visueller Aufmerksamkeit in den Vordergrund gestellt haben und ein zweites Anwendungsexperiment auf einer allgemeinen Objekterkennungsdatenbank hinzugefügt haben. Diesen Beitrag haben wir schlussendlich bei der Internationalen Conference on Computer Vision and Pattern Recognition (CVPR) 2020 eingereicht. Wir erhoffen uns eine positive Teilnahme, die erhaltenen Reviews für die ICCV Teilnahme waren jedenfalls sehr hilfreich. Im Moment läuft die Begutachtung bei der CVPR, aber der Beitrag ist bereits unter als ArXiv Dokument verfügbar [20].

Chip-Prototyp basierte Fehlererkennung, via biologischen Gehirnmodellen und Gabor-Filter

Da das Chip-Datenmaterial so heterogen sein kann, was sich in den vorhergehenden Arbeiten in Jahr 4 gezeigt hat, explorieren wir in Zusammenarbeit mit dem Stifter neue Ansätze aus. Idee: Einen Prototypen eines Chips zu erzeugen, und damit gegen andere Chips zu

[20] Titel: „Combining a Biologically-Plausible Model of Visual Attention with Deep Learning to Improve Wafer Fault Detection“

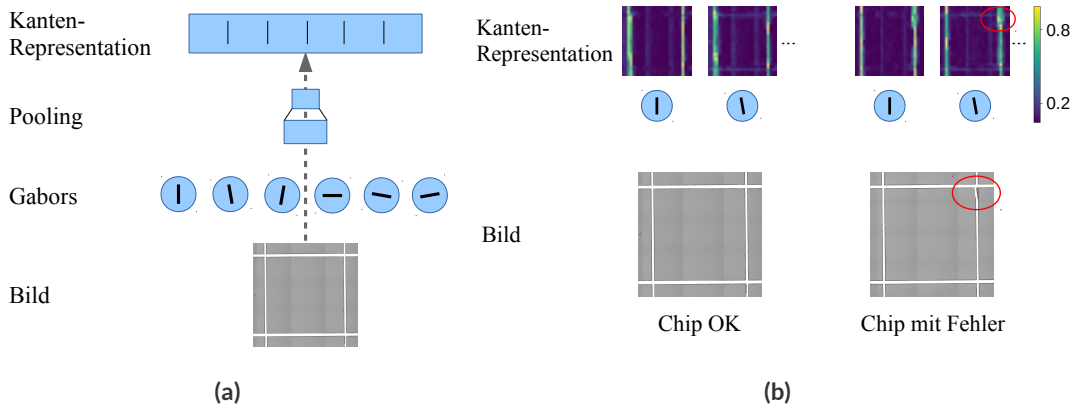


Abbildung 67: a) Geplantes Modell. **b)** Vorläufige Ergebnisse: Gaborfilter und Antwort.

testen. Das Erzeugen eines Prototypen umgeht schön das Problem der Bildheterogenität. Die Idee basiert auf den Stifer 3D Micromac. Die Erzeugung selbst geht mit geringen manuellen Aufwand und es damit auch für einen späteren Prozessingenieur in der Fertigung möglich. Das ursprüngliche Problem war, dass das Convolutional Neuronal Network (CNN) zwar schon besser als alle anderen Forschungsansätze war wie sich im ICCV/CVPR- und ETFA-Beitrag gezeigt hat, aber immer noch die Fehler zu 'nur' 90% erkannt wurden. Diese *Accuracy* ist noch etwas wenig um in einem Industrieprozess sehr gut eingesetzt werden zu können.

Ein Chip-Prototypen auf Bildebene funktioniert nicht so gut, da es jede Menge Bildartefakte gibt (Pixelrauschen, variable Anteile, Verzerrungen und Helligkeitsprobleme). Deswegen versuchen wir erst einmal eine höhere bzw. abstrakte Repräsentation des Bildes zu erschaffen, wo die Kanten gut repräsentiert werden (diese sind das Wichtige). Dies hat gewisse Ähnlichkeiten mit den Arbeiten von Frederik Beuth über die ersten Areale im visuellen System des Menschen (primärer visueller Kortex, V1) [26]. Der primäre visuelle Kortex filtert das Bild u.a. nach orientierten Kanten, welche mittels Gabor-Filter modelliert werden können [61]. Deswegen versuchen wir nun aus zu testen in wie weit sich biologische Verarbeitungsprinzipien übernehmen lassen und die Erkennungskraft des menschlichen Gehirnes übertragen lässt.

Die Arbeit ist noch nicht komplett abgeschlossen, aber das Modell ist in etwa das Folgende (Abbildung 67a). Das Eingabebild wird in eine Kanten-Repräsentation überführt, dann die räumliche Auflösung verringert und Kontraste verstärkt. Die abschließende Repräsentation kann dann als Chip Prototyp benutzt werden, und versus andere Chips verglichen werden. Dazu soll dann OpenCV-Template Match auf dieser Repräsentation ausgeführt werden. Die aktuellen Resultate sind vorläufig (Abbildung 67b), zeigen aber schon die Mächtigkeit des neuen Ansatzes an.

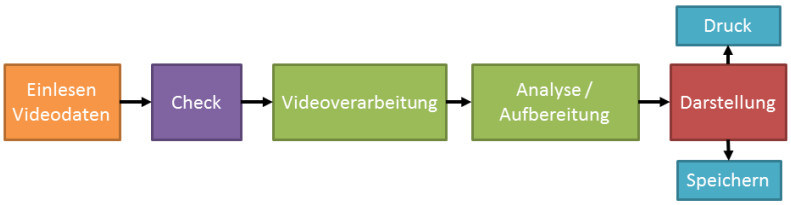


Abbildung 68: Ablauf des Programmablaufs

Dieses Projekt wird, mit zwei leicht verschiedenen Ausrichtungen, im Rahmen zweier studentischer Arbeiten durchgeführt:

- W. Farooq, "Edge detection and Contrast Enhancement of Wafers using Collinearity," Master Thesis, TU Chemnitz, Chemnitz, 2019.
- M. Friedrich, "Binäres Template Matching auf Gabor Bildern zur Fehleranalyse in der Halbleiterindustrie," Forschungspraktikum, TU Chemnitz, Chemnitz, 2019.

2.5.5 Restliche Zuarbeiten für den Stifter - Schnittstellendefinition und grafische Nutzeroberflächen

Schmelzzonen-Analyse Die visualisierten Ergebnisse können gedruckt oder auch in einer Datei gespeichert werden um bei einem späteren Laden des Datenstamms den Bearbeitungsprozess nicht erneut starten zu müssen. Bei der Umsetzung dieses PC Werkzeuges und dessen Abarbeitungsschritte (siehe Bild 68) wurde auf Modularität und Wiederverwendbarkeit der einzelnen Programmteile geachtet. Des Weiteren wurden auch die *Software Patterns*, die im ersten Berichtszeitraum recherchiert wurden, in dieser Entwicklung auch mit angewendet. Das im Rahmen des Konferenzbeitrags von John *et al.* ([59]) entstandene Anwenderprogramm ist frei erhältlich für die Community, zu finden unter dem Name Weldpool Evaluation Tool auf der Projektwebseite[21].

Wafer-Analyse

In der ersten Phase wurde ein regelbasierte Ansatz zur Bildanalyse entwickelt und in ein neu entwickeltes Programm namens „Wafer Analysis Tool" integriert. Dessen grafische Nutzeroberfläche erlaubt das Lesen großer Wafer-Bilder und die Definition einer elliptischen ROI. Letztere wird initial automatisch an der Außenkante des Wafers erkannt. Danach werden die einzelnen Chips per Kantendetektion erkannt. Sie werden in der Darstellung als Rechtecke farbig markiert und werden zeilen- und spaltenweise von -12 bis +12 laufend adressiert. Die am Rand liegenden unvollständigen Chips werden mit der Farbe oran-

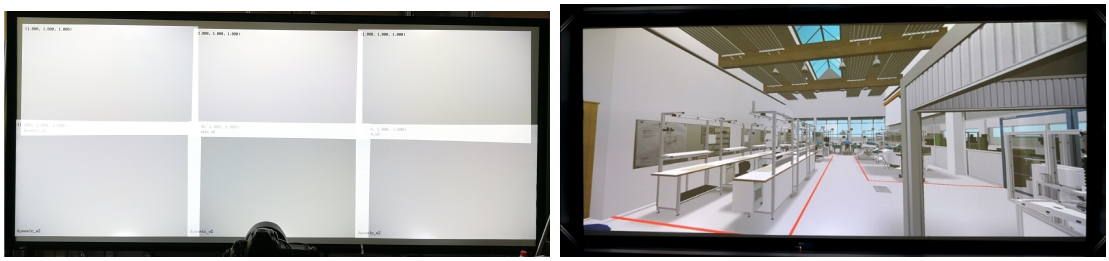
[21] `https://localize-it.de/downloads/`

ge markiert und basieren auf dem Vorwissen der Rand-Region. Chips, die übersehen werden vom zuvor genannten Verfahren werden interpoliert und blau markiert. Schädigungen, basierend auf einer Histogramm-Analyse werden rot markiert. Das Programm zeigt dann zunächst die vermeintlich geschädigten Chips an, die dann vom Nutzer per *Pushbutton*-Eingabe schnell bestätigt oder falsifiziert werden kann. Dadurch ist es möglich eine *Ground Truth* zu erstellen. Die generierten Ergebnisse werden kommasepariert als Tabelle ausgeben und können so innerhalb des Klassifikationsverfahrens zur Parameteroptimierung benutzt werden. Damit ist es nachher auch möglich einen Klassifikator zu trainieren oder das vorhandene Verfahren mit seinen Parametern zu optimieren. Dank Code-Optimierung und Parallelisierung ist ein Wafer auf einer CPU mit 8 Kernen bereits in wenigen Minuten prozessiert.

2.5.6 Bilanz nach Erreichen des letzten Meilensteins in AB 4

In dem AB 4 wurde der Transfer von Deep Learning Knowledge von der mehr in der Objekterkennung gelegenen Domain in die Industriedomain erfolgreich durchgeführt. Dies ist eine natürliche Bedingung für die lokale Industrie, die um Chemnitz herum traditionell im Bereich Maschinenbau sehr stark aufgestellt ist, mit Unternehmen die teilweise Weltmarktführer sind auf ihren Bereichen sind ('Hidden Champions'). In unserem Fall war es der Stifter und Kooperationspartner 3D Micromag AG, im Bereich Laser Mikrobearbeitung. In den ersten beiden Projektjahren wurde das Thema in situ-Charakterisierung von Schmelzprozessen erfolgreich durch geführt und umfänglich evaluiert, was schönerweise in zwei Publikationen auf einer nationalen und einer internationalen Konferenzen resultiert hat. Anschließend hatte sich gezeigt, dass das Thema schon ausreichend performant funktioniert hat, und das es keine wirtschaftliche Notwendigkeit für eine Steuerung und Kopplung gab, so dass anstelle der Antrag vorgesehen Themen auf Wunsch des Stifters ein neues Thema eruiert werden sollte, die Thermal Laser Separation (TLS). Bei ihr werden Halbleiterwafer mit neuartigen Laserschneidprozessen in einzelne Chips getrennt. In dem Bereich ist die Herausforderung - neben dem Transferproblem - dass die zuerkennenden Fehler sehr klein sind und das Bildmaterial recht heterogen ist. Deswegen wurde ein neuartiger Ansatz entwickelt, basierend auf Deep Learning und biologisch inspirierter Verarbeitung, welcher das Problem löst. Da dieser Ansatz sehr neu ist, bietet er nicht nur für die Maschinenbaudomain Potential, sondern ist auch für die Computervision-Community. Er wurde probeweise auf den Computervision-A-Level Konferenzen International Conference on Computer Vision (ICCV) und Conference on Computer Vision and Pattern Recognition (CVPR) eingereicht, bei der letzteren läuft die Begutachtung noch. Die Evaluation für die letztere Konferenz hat gezeigt, dass wir vermutlich das international am besten funktionierende System für die Domain entwickelt haben. Wir sagen vermutlich, da das System besser als jeder veröffent-

lichter State-of-the-Art abschneidet, aber oft in der Domain nicht alle Verfahren publizierte werden. Parallel dazu wurde das System für die Industrie-Community aufbereitet und verwertet, was einmal eine Veröffentlichung auf dem wissenschaftlichen Projektworkshops bei den Chemnitzer Linux-Tagen 2019 nach sich zog, und auf einer internationalen Maschinenbaukonferenz resultiert ist. Von seitens des Stifters wurde großes Interesse bekundet das Thema fortzusetzen.



(a) (b)

Abbildung 69: Ausgangssituation nach dem ersten Einschalten eines Multiprojektorsystems (a). Nach erfolgreicher Kalibrierung wirkt die Projektionsfläche wie ein einheitliches Bild (b).

2.6 AB 5: Deviceless 3D-Steuerung

Der Arbeitsbereich befasst sich mit der Kalibrierung, Ansteuerung und Interaktion mit Multiprojektorsystemen (siehe Abbildung 69). Die Kalibrierung dieser Systeme unterteilt sich in zwei Teilprobleme. Die geometrische Kalibrierung erfasst die tatsächliche Bildgeometrie sowie Lage und Orientierung jedes Projektors. Ist die Position jedes Projektorpixels bezüglich der Leinwand erfasst, so müssen anzuzeigende Bilddaten entsprechend der Projektoranordnung aufgeteilt und invers verzerrt werden, so dass auf der Leinwand das unverzerrte Gesamtbild erscheint. Neben der geometrischen Kalibrierung ist auch die photometrische Kalibrierung von entscheidender Bedeutung. Ziel ist es, dafür zu sorgen, dass ein Betrachter das gesamte Display als einheitlich Hell und mit identischer Farbwiedergabe wahrnimmt, so dass der segmentierte Charakter überhaupt nicht mehr erkennbar ist. Die Herausforderung besteht darin den Kalibriervorgang, also die Ermittlung des Ist-Zustands, vollständig zu automatisieren, wobei die Projektorausgaben mittels einer oder mehreren Kameras erfasst werden.

Die Effizienz der Bildverarbeitung stellt einen weiteren wichtigen Aspekt sowohl bei der geometrischen als auch bei der photometrischen Kalibrierung dar. Zielstellung ist der Betrieb großflächiger, hochauflösender Displays. Beim Einsatz von 6 FullHD-Projektoren ergeben sich über 12 Megapixel, die im Idealfall 60 mal pro Sekunde verarbeitet werden müssen. Im stereoskopischen Betrieb verdoppelt sich die Datenmenge nochmals. Daher ist es unerlässlich, Verfahren zu entwickeln, die die parallelen Berechnungseinheiten moderner GPUs ausnutzen.

In der letzten Phase des Projekts wird untersucht, inwieweit sich die Ergebnisse aus AB1 integriert lassen, um kamerabasiert den Anwender zu erkennen und ihm so die Steuerung von Anwendungen auf dem Display zu ermöglichen. Die Anlage kann dann vollständig ohne Hilfsmittel bedient werden.

2.6.1 Aufgabenstellung und Zielsetzung

Im Projekt wurden Aufgaben quartalsweise formuliert und auf diese Weise auch in den Zwischenberichten abgehandelt. In der folgenden Auflistung sollen daher die übergeordneten Inhalte betrachtet und die Lösung kurz angerissen werden:

1. Kompensation der geometrischen und photometrischen Verzerrung von Aufnahmegeräten:
   Für die Lokalisierung geometrischer Verzerrungen wurde ein Verfahren entwickelt, bei dem die Form der Leinwand selbst als Referenz dient. Für die Berechnung der photometrischen Verzerrung wurde ein Teststand entwickelt, um die Abbildung des Kamera-Farbraumes in einen geräteunabhängigen XYZ-Farbraum zu berechnen. Eine Erweiterung des Verfahrens ermöglicht die direkte Abbildung zwischen Farbräumen verschiedener Kameras.

2. Kompensation der geometrischen und photometrischen Verzerrung von Projektoren:
   Für die geometrische Kalibrierung mehrerer Projektoren wurde ein Verfahren, basierend auf structured light pattern, entwickelt. Neben dem Ausgleich der Trapezverzerrung berechnet das Verfahren auch ein Mapping zwischen dem Wall-Space und dem Projektor-Space. Für die photometrische Kalibrierung wurde ein Verfahren entwickelt, dass die Transferfunktion zwischen dem Projektor und dem RGB-Farbraum berechnet. Die Schwierigkeit hierbei lag in der Anpassung der Verfahren auf verschiedene Farbräder sowie bei der Kompensation der daraus resultierenden zeitsequentiellen Farberzeugung.

3. Echtzeitfähige Verarbeitung mehrerer Eingangssignale:
   Für die Aufbereitung der darzustellenden Bilddaten wurden vier verschiedene Software-Ebenen identifiziert, an denen die entwickelte Kalibrier-Bibliothek integriert werden kann.

4. Erfassung von Nutzerposition und -eingaben:
   Zur Bestimmung der Position sowie der Pose des Nutzers wurde auf Methoden des maschinellen Lernens zurückgegriffen. In einem Versuchsaufbau konnte anhand eines interaktiven Schachspiels die Eignung der verwendeten Methoden untersucht werden.

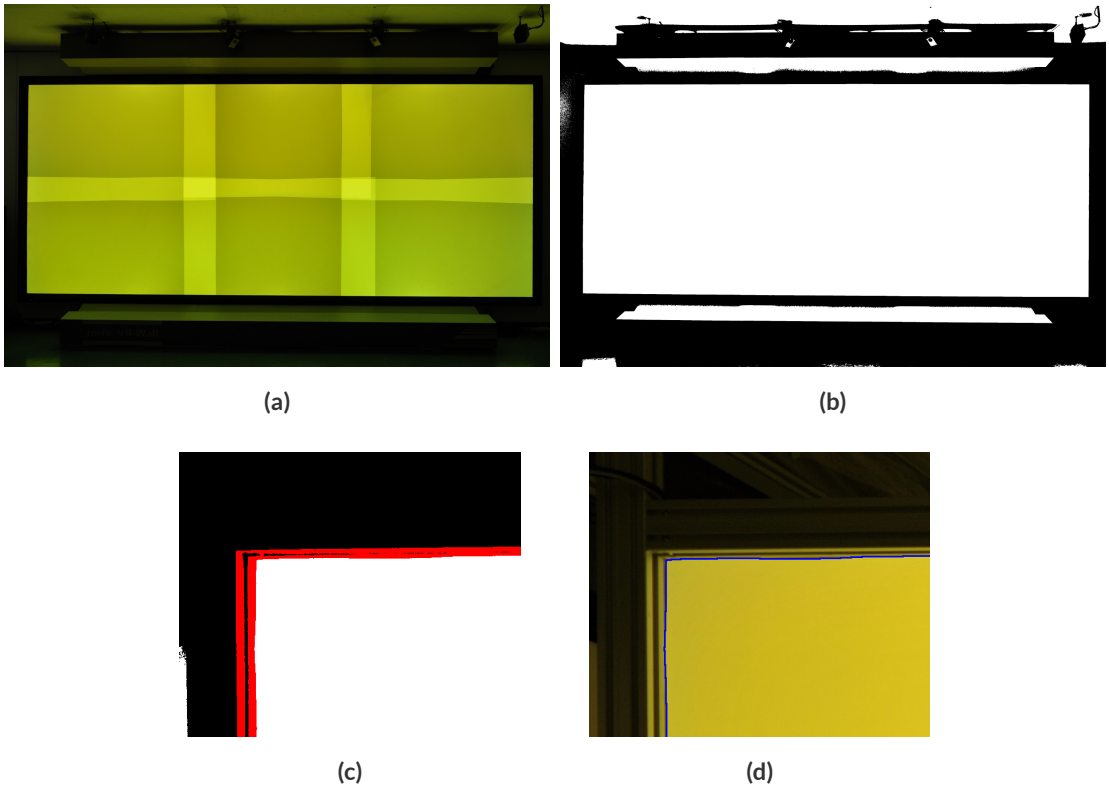


Abbildung 70: Geometrische Kalibrierung der Aufnahmegeräte unter Verwendung der Geometrie der Leinwand. In einem ersten Schritt wird ein Bild der kompletten Leinwand aufgezeichnet (a) und binarisiert (b). Inerhalb der s/w-Darstellung wird mit Hilfe von Methoden der Bildverarbeitung die Randregion der Leinwand gesucht (c). Ein Subpixel-genaues Schätzverfahren liefert den Grenzverlauf der Leinwand (d).

2.6.2 Kompensation der geometrischen und photometrischen Verzerrungen bei Aufnahmegeräten

Für die Kompensation der geometrischen Verzerrung wurde ein Kalibrierverfahren entwickelt, welches sich die speziellen Gegebenheiten bei der Kalibrierung von Projektionsanlagen zu nutze macht: Die Form der Leinwand selbst kann als Referenz zur Kalibrierung herangezogen werden. Dazu muss die Leinwand fotografisch erfasst und Begrenzung der Leinwand im Foto detektiert und die Verzerrung über die gesamte Leinwand interpoliert werden (siehe Abbildung 70). Da ein Kamerapixel im Regelfall mehreren Projektorpixeln entspricht und unstetige Sprünge über mehrere Projektorpixel zu sichtbaren Artefakten führen würden, ist eine Subpixel-genaue Schätzung des Grenzverlaufs notwendig.

Im Gegensatz zu allgemeinen Verfahren zur geometrischen Kalibrierung von Kameras gelten die ermittelten Korrekturparameter nur für die tatsächliche Aufnahmeposition und

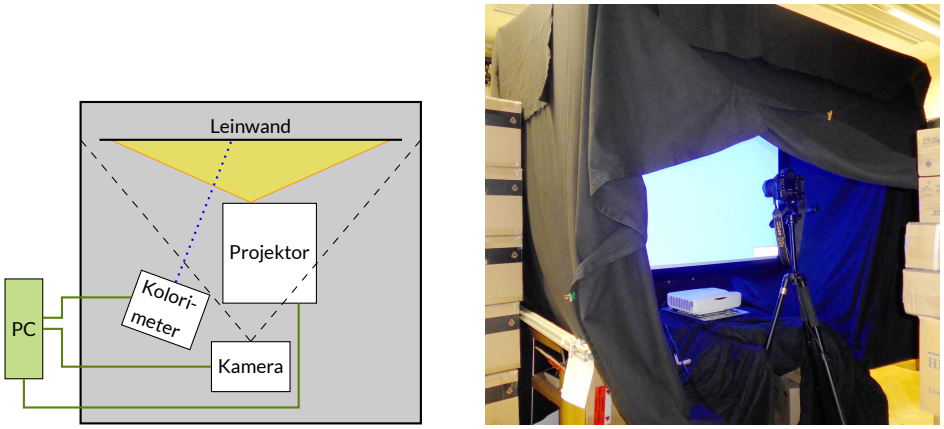


Abbildung 71: Aufbau des verwendeten Teststandes zur photometrischen Kalibrierung von Aufnahmegeräten

können insbesondere nicht für andere Entfernungen und Winkel wiederverwendet werden. Da allerdings nur ein einziges Foto notwendig ist, lässt sich das Verfahren direkt in den Workflow zur geometrischen Kalibrierung integrieren, so dass eine Vorkalibrierung des Aufnahmegerätes generell entfallen kann.

Photometrische Verzerrungen bei Aufnahmegeräten äußern sich durch Unterschiede im aufgenommenen Farbwert und der Helligkeit. Zur Erkennung der Abweichungen ist eine Analyse der Farberfassungseigenschaften der Kamera notwendig. Gängige Verfahren verwenden dazu Farb-Testcharts und eine hochwertige Referenzlichtquelle, die jedoch bei einem späteren Einsatz der Verfahren nicht zur Verfügung stehen.

Um dennoch eine Erfassung zu ermöglichen, wurde ein Verfahren entwickelt, bei dem die ein Projektor als Lichtquelle benutzt wird. Ein blickdicht verschließbarer Teststand für Ultrakurzdistanz-Projektoren ermöglicht dabei Messungen mit konstanten Umgebungsbedingungen (Abbildung 71).

Durch Messung einer Vielzahl von Farbsamples sowohl mit der zu analysierenden Kamera als auch mit einem hochwertigen Spektralphotometer ließen sich die einzelnen Messwerte im Kamera-Farbraum mit den Referenzwerten im geräteunabhängigen XYZ-Farbraum in Beziehung setzen und so die Abweichungen bestimmen und über den gesamten Farbraum charakterisieren (siehe Abbildung 72). Diese Vorgehensweise hat lediglich den Nachteil, dass sie keine direkten Aussagen außerhalb des Gamuts der eingesetzten Projektoren erlaubt, was sich jedoch für den geplanten Einsatz nicht als hinderlich erweisen sollte.

Durch die Erfassung eines möglichst regelmäßigen Gitters bezüglich des RGB-Eingabefarbraums der Projektors mit jeweils einer Referenzmessung als auch mit der zu vermessenden Kamera können die Gitterzellen im Farbraum aufeinander abgebildet werden, und es lässt sich innerhalb jeder Gitterzelle eine stückweise lineare Interpolation

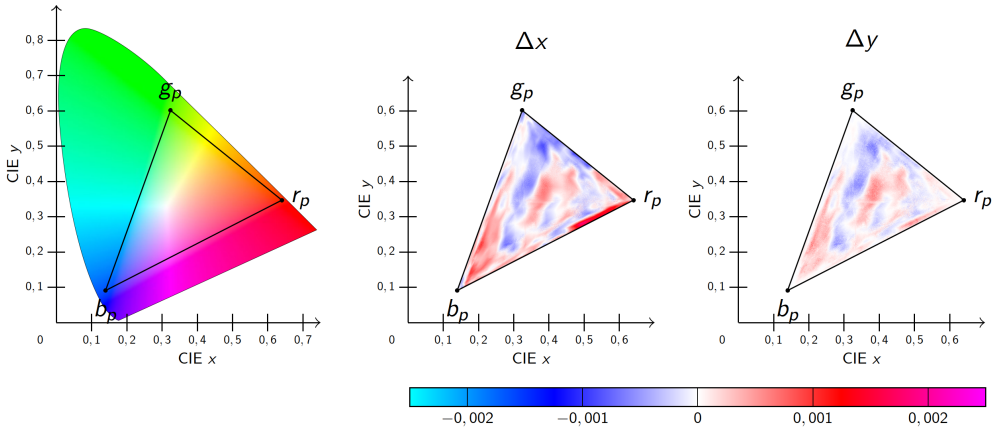


Abbildung 72: Visualisierung der Kamera-Abweichungen in einem Slice bei $Y = 0,7$ im CIE Chromazitätsdiagramm auf Basis von 4096 Samples. Aufgezeichnet mit einer Nikon D3s unter Verwendung eines Sanyo PDG DWL-2500 Projektors.

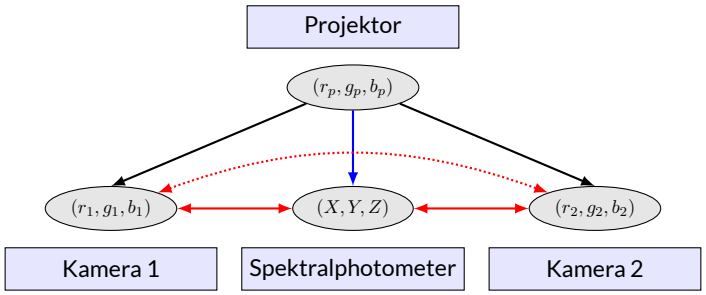


Abbildung 73: Farbkalibrierung der Kameras: Durch Messung der gleichen Farbsamples (angeordnet in einem Gitter bezüglich des RGB-Frabraums des Projektors, der das Pattern generiert) mit mehreren Kameras sowie einem Spektralphotometer als Referenz lassen sich Abbildungen der Farbräume jeder Kamera in den geräteunabhängigen XYZ-Farbraum - und schließlich auch direkt zwischen den Kameras - erstellen.

basierend auf Tetraedern vornehmen. Dies ermöglicht es, eine Abschätzung der Charakteristik der Farberfassung einer Kamera über die tatsächlich gemessenen Gitterpunkte hinaus für den gesamten Gamut des Projektors zu erstellen. Eine derartige Vorgehensweise ist besonders geeignet für den Ausgleich von Farbabweichungen zwischen verschiedenen Kameras.

Durch Erfassung der Farbsamples mittels eines hochwertigen Spektralphotometers können Referenzwerte bestimmt und mittels des Interpolationsverfahrens zwischen den Messergebnissen unterschiedlicher Kameras und dem Referenz-Farbraum transformiert werden. Von dort aus können die Ergebnisse durch die inverse Abbildung auch in den nativen Farbraum jeder anderen Kamera rücktransformiert werden, so dass sich auch die

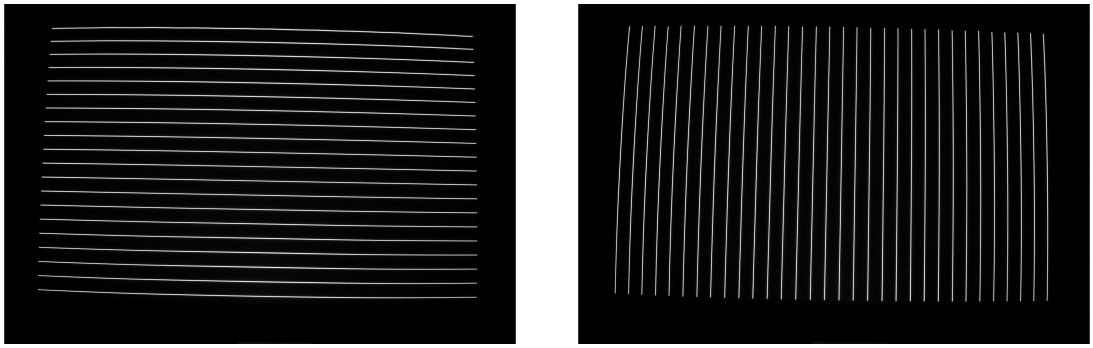


Abbildung 74: Vertikales und horizontales structured light pattern zur Berechnung der geometrischen Verzerrung eines Projektors.

Messwerte einer Kamera in die erwarteten Messwerte, die eine andere Kamera liefern würde, umrechnen lassen, wie in Abbildung 73 veranschaulicht. Sofern zur Einmessung mehrerer Kameras der gleiche Projektor verwendet wurde, können deren Ausgaben über den gemeinsamen Projektor-Farbraum direkt zueinander in Beziehung gesetzt werden. Dies ermöglicht es auch perspektivisch, die Einmessung der Kameras später direkt vor Ort durchführen zu können, ohne das dafür auf teures Spezialequipment wie Spektralphotometer zurückgegriffen werden muss.

2.6.3 Kompensation der geometrischen und photometrischen Verzerrung von Projektoren

Der Stiftungspartner 3DInsight ist insbesondere am Einsatz von ultra-weitwinkligen Projektoren interessiert, welche bauartbedingt deutliche Verzeichnungen verursachen. Weiterhin führt die geringe Projektionsentfernung dazu, dass bereits geringe Unebenheiten der Projektionsfläche zu sichtbaren Verzerrungen führen. Zur Lösung des Problems wurde ein Verfahren basierend auf structured light pattern implementiert. Das Verfahren berechnet in zwei Schritten eine Registrierung zwischen Kamera- und Projektorbildraum anhand von Punktgittern.

Im ersten Schritt wird die Verzerrung jedes einzelnen Beamers unter Verwendung von horizontalen und vertikalen Pattern berechnet (siehe Abbildung 74). Anschließend erfolgt die Berechnung der Verschiebungen zwischen benachbarten Segmenten der Projektionswand. Hierzu wird ein Gitter-pattern verwendet, was in Abbildung 76 dargestellt ist. Das Ergebnis der geometrischen Kalibrierung zeigt Abbildung 75.

Bei den Experimenten trat ein weiteres Problem auf, welches bisher in der Forschung noch nicht untersucht wurde: Thermische Effekte führen zu einer Veränderungen der relativen Position zwischen Projektor und Leinwand. Vor allem bei den verwendeten extrem weitwinkligen Projektoren führt das zu wahrnehmbaren Abweichungen. Im Rahmen das Pro-

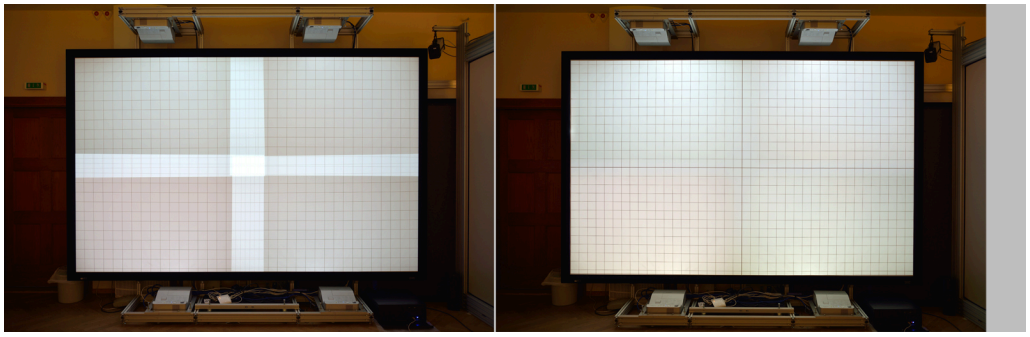


Abbildung 75: Vergleich zwischen nicht kalibrierten Projektionssegmenten (links) und der Darstellung des verwendeten Gitters nach der geometrischen Kalibrierung (rechts).

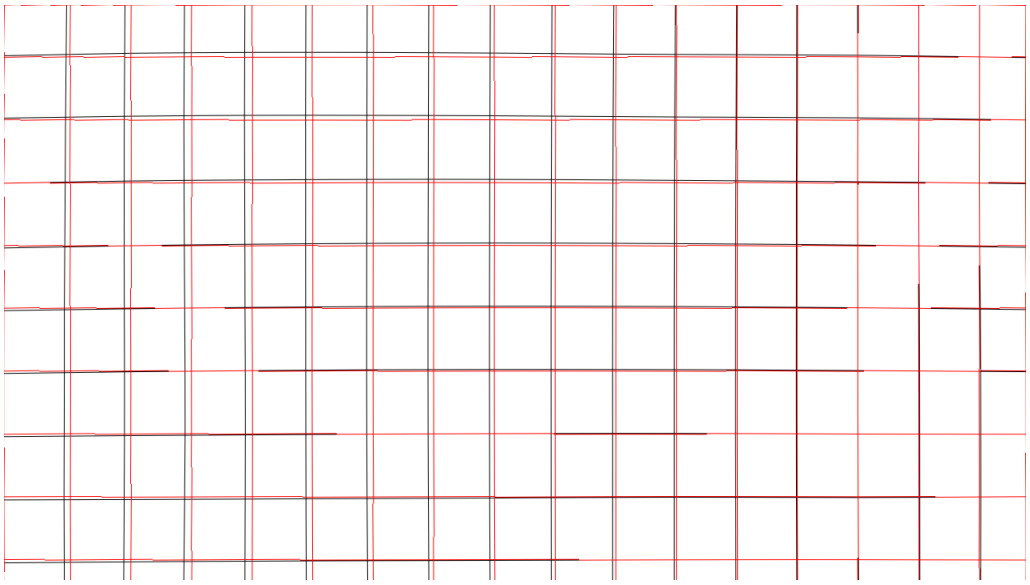


Abbildung 76: Darstellung der geometrischen Verschiebungen zwischen zwei Segmenten einer Projektionswand unter Verwendung eines Gitter-pattern.

jekts wurde deshalb ein Verfahren entwickelt, um möglichst effizient mit dieser Situation umgehen zu können. Unter der Annahme, dass sich weder die Eigenschaften der Optik noch die Form der Leinwand verändern und die auftretenden Verschiebungen nur gering sind, lässt sich damit ein eingemessener Kalibrierdatensatz auf eine veränderte Position übertragen. Dazu ist lediglich die aktualisierte Position der Eckpunkte des Projektorbildes erforderlich, welche sich mit einem einzigen Foto erfassen lässt, so dass eine sehr schnelle Korrektur erfolgen kann und eine komplette Neukalibrierung vermieden wird.

Da dieser einfache Ansatz nicht in allen Fällen ausreichen, wurde ein das Kalibrierverfahren erweitert. Dazu wurden in Abhängigkeit von der Temperatur mehrere Datensätze aufgezeichnet. Ausgehend von den gemessenen Kalibrierdatensätzen lässt sich durch Interpolation entsprechend der gemessenen Temperatur während des Betriebs der Powerwall

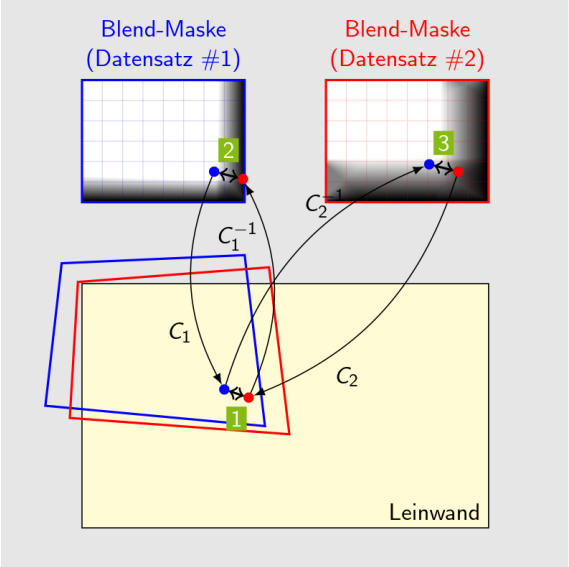


Abbildung 77: Die Leinwand-Position für ein gegebenes Pixel wird durch Interpolation der Bildraum-Positionen aus zwei Kalibrierdatensätzen berechnet.

eine individuelle Blend-Maske berechnen. Abbildung 77 zeigt schematisch das entwickelte Interpolationsverfahren. Das Verfahren wurde vollständig umgesetzt und funktioniert mit synthetischen Datensätzen. Ob es sinnvoll ist ein oder mehrere Temperatursensoren zu verwenden und wie viele Datensätze notwendig sind, müssen Praxistests des Stifterunternehmens 3DInsight zeigen.

Das entwickelte Verfahren wurde für den Einsatz von mehreren Aufnahmegeräten erweitert, wofür Verfahren zur Berechnung eines gemeinsamen Parameterraums entwickelt wurden. Nach Rücksprache mit dem Stiftungspartner 3DInsight konnten drei relevante Szenarien identifiziert werden.

1. Die komplette Bildfläche der Projektion wird von allen Kameras erfasst:
   In diesem Fall kann die Geometrie der Leinwand direkt als gemeinsamer Parameterraum verwendet werden, so dass sich die 2D-Bildraumkoordinaten zwischen den einzelnen Kameras und auch der Leinwand umrechnen lassen.

2. Verwendung einer zusätzlichen Übersichtskamera, die die komplette Leinwand erfasst:
   Wenn nicht alle Kameras die komplette Leinwand erfassen, kann mit einer höheren Auflösung auf den unterschiedlichen Teilsegmenten gearbeitet werden kann wodurch sich das Kalibrierergebnis verbessert. Über die einzelnen geometrischen Kalibrierpattern der Projektoren lässt sich für jede Kamera das Mapping in den gemeinsamen Überblicks-Bildraum bestimmen, sofern die Kamera mindestens ein ganzes Projektor-Segment erfasst. An das Überblicksbild werden dabei keine hohen Anfor-

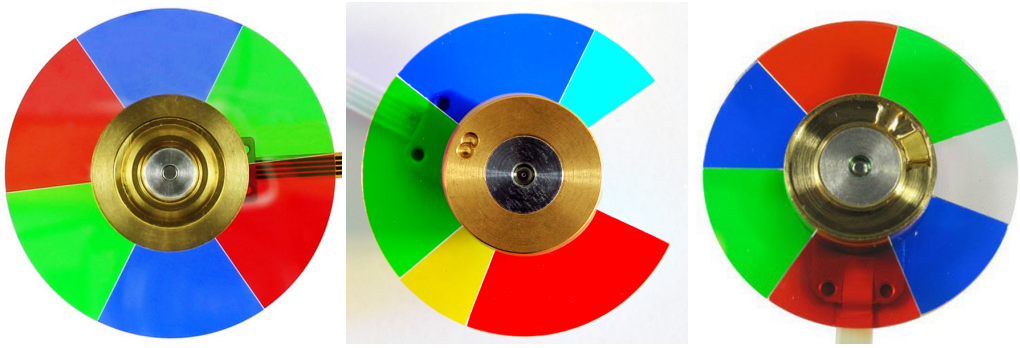


Abbildung 78: Darstellung verschiedener in DLP-Projektoren typischerweise vorkommender Farbräder.

derungen bezüglich Auflösung oder Genauigkeit gestellt, sondern es dient nur zur groben Ausrichtung der Einzelkameras.

3. Verwendung von Nachbarschaftsbeziehungen zwischen den Kameras:
   Das Verfahren setzt voraus, dass es paarweise zwischen den Kameras immer einen Überlappungsbereich von mindestens einem Projektor-Segment gibt. Dadurch lassen sich benachbarte Kameras über das gemeinsame Segment registrieren. Da sich bei der Korrektur der geometrischen Kameraverzerrungen Fehler nie komplett reduzieren lassen, muss ein weiterer Optimierungsschritt erfolgen, bei dem diejenigen Kamera-Verzerrungsparameter für zwei Kameras gesucht werden, die die Abstände zwischen korrespondierenden Feature-Punkten des gemeinsamen Projektor-Segments minimieren.

Ein weiterer Arbeitspunkt war die Farbkalibrierung der Projektoren. Das Stifterunternehmen 3DInsight GmbH war dabei insbesondere an den spezifischen Charakteristiken von DLP-Projektoren interessiert. Die kamerabasierte Erfassung der Projektorausgaben wird dabei durch die *zeitsequentielle* Farberzeugung deutlich erschwert.

Die Farbkanäle (rot, grün, blau, ggf. noch weitere - siehe Abbildung 78) werden nicht gleichzeitig projiziert (wie z.B. bei LCD-Projektoren), sondern ein rotierendes Farbrad im Strahlengang sorgt dafür, dass zu jedem Zeitpunkt nur ein Farbkanal angezeigt wird, und das Gesamtbild durch die zeitlich versetzte Darstellung aller Kanäle erst durch die Trägheit des menschlichen Auges entsteht. Bei der Erfassung mit Kameras ergibt sich dann das Problem, dass die Kamera nicht mit dem Farbrad synchronisiert ist, und es zu systematischen Messfehlern kommt, wenn die Belichtung nicht exakte ganze Farbrad-Umdrehungen enthält. Der Fehler wird dabei umso größer, je kürzer die Belichtungszeit gewählt wird, wohingegen kurze Belichtungszeiten im Hinblick auf die gesamte Messdauer besonders wünschenswert sind. In Abbildung 79 ist der Sachverhalt dargestellt.

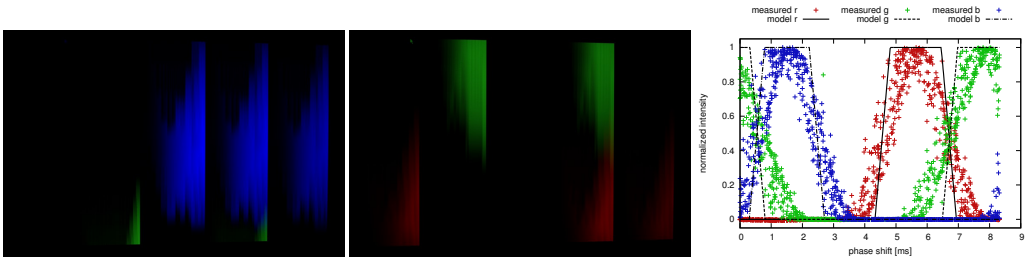


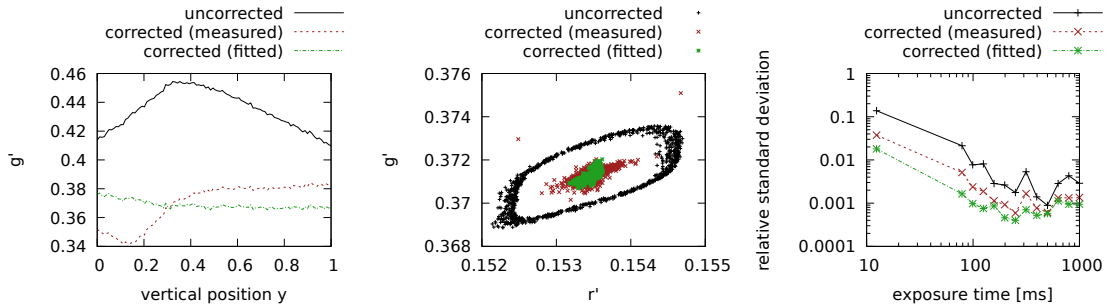
Abbildung 79: Fotographische Erfassung der Projektionsfläche mit Belichtungszeit 1/2000s (links und mitte). Darstellung der Korrelation zwischen Auslösezeitpunkt und gemessenem Farbwert (rechts).

Eine Synchronisation auf Seiten der Kamera ist nicht möglich, da die meisten Kameras nach Empfang des Kommandos eine variable, nicht vorhersehbare Auslöseverzögerung in der Größenordnung von 5 bis 100 Millisekunden aufweisen. Es ist jedoch möglich den exakten Auslösezeitpunkt nachzuvollziehen und dadurch entstandene Fehler zu kompensieren. Da die Bilderzeugung des Projektors - und damit die Position des Farbrades - mit der Grafikkarte des Rechners synchronisiert ist, lässt der gemessene Auslösezeitpunkt in Relation zum letzten Bildwechsel-Zeitpunkt der Grafikkarte Rückschlüsse auf die Position des Farbrads während der Aufnahme zu.

Abbildung 80 zeigt die Ergebnisse des Korrekturverfahrens anhand von 1000 Aufnahmen des gleichen Projektorbildes. Dargestellt ist die Projektion der R,G,B-Messwerte auf die von den Grundfarben der Kamera definierte 2D-Farbebene (die Helligkeit wird dabei ignoriert). Jeder Punkt dieser Ebene repräsentiert einen anderen Farbton, der Urpsrung $(0; 0)$ entspricht dem maximal gesättigten Blau, $(1; 0)$ dem maximal gesättigten rot und $(0; 1)$ dem maximal gesättigten Grün, und der gesamte von der Kamera erfasste Farbraum liegt innerhalb dieses Dreiecks. Bei der Aufnahme der Weißphotos des Projektors sollte sich der gleiche Punkt für jedes der Fotos ergeben. Da jedes Foto das Farbrad jedoch mit einem anderen Phasenversatz erfasst hat, ergibt sich ein systematischer Messfehler in Form eines verzerrten Pfades um das eigentliche Ergebnis herum - ein einzelnes Foto liefert also zunächst nur einen beliebigen Punkt auf diesem Pfad. Das entwickelte Korrekturverfahren kann die Resultate deutlich verbessern, die Messwerte liegen nun normalverteilt um den tatsächlichen Weißpunkt des Projektors.

2.6.4 Echtzeitfähige Verarbeitung mehrerer Eingangssignale

Nach Rücksprache mit dem Stiftungspartner 3DInsight wurden vier verschiedene Software-Ebenen identifiziert, wie die entwickelten Verfahren in bestehende Anwendungen integriert werden können (siehe Abbildung 81):



(a) Korrektur für $\Delta_e \approx 0.0125s$ **(b)** Chromatizität für $\Delta_e \approx 0.4s$ **(c)** Vergleich verschiedener Δ_e

Abbildung 80: Ergebnisse der Korrektur: (a) Effekt der Korrektur über eine vertikale Pixelzeile des Projektors. (b) Chromatizitätskoordinaten von 1000 Fotos eines Projektors, mit konstantem Farbwert. Die Rohdaten liegen auf einem Pfad um den tatsächlichen Wert. Die Anwendung unserer Korrekturmethode führt zu einer eher gaußartigen Verteilung um den Mittelwert. (c) Vergleich der relativen Standardabweichung in der Chromazitätskoordinate $g\prime$ bei verschiedenen Expositionszeiten (Log-Log-Diagramm).

1. Bibliothek innerhalb der Anwendung:
   Eine Programmbibliothek mit wohldefinierter Schnittstelle nimmt die Bilddaten einer graphischen Anwendung entgegen und führt die für die geometrische und photometrische Kalibrierung notwendigen Nachverarbeitungsschritte durch. Zur Nutzung der Bibliothek muss die Anwendung derart modifiziert werden, dass statt des normalen Framebuffers in eine Textur gerendert wird. Die Kalibrier-Bibliothek übernimmt dann das Rendern dieser Textur in den tatsächlichen Framebuffer, unter Anwendung der entsprechenden Kalibrierverfahren. Die Bilddaten verlassen dabei niemals die GPU, so dass die Darstellungslatenzen minimiert werden.

2. Umsetzung als Kompnente einer Render-Engine:
   Die Bibliothek wird dabei als Plugin in eine Render-Engine integriert. Dadurch können potentiell alle auf dieser Engine basierenden Anwendungen direkt ohne weitere Anpassungen die Projektionsanlage nutzen. Beispielhaft erfolgte eine Integration in die Unreal Engine 4.

3. Integration in das Betriebssystem:
   Bei der Umsetzung erwies sich die Tatsache als hilfreich, dass praktisch alle modernen graphischen Oberflächen sogenanntes *Desktop Compositing* umsetzen, bei denen die Ausgabe der einzelnen Programme als Texturen auf der GPU vorliegen. Unter Verwendung einer OpenGL-Erweiterung kann auf den Inhalt des Windows-Desktops als Textur zugegriffen werden und durch eine andere Textur ersetzt werden. Aufbauend

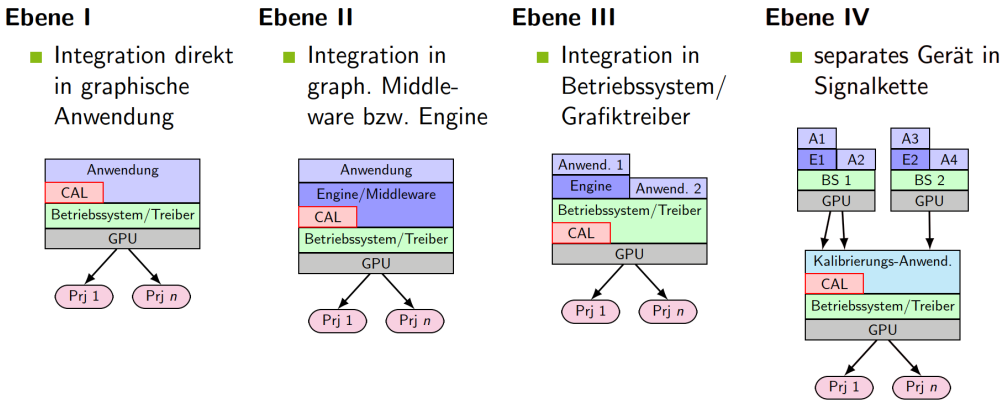


Abbildung 81: Darstellung der identifizierten Ebenen für die Anwendung der Kalibriermethoden (im Bild als Rechteck mit der Bezeichnung CAL dargestellt).

darauf wurde eine Software entwickelt, die die Kalibrierung direkt auf den Windows-Desktop anwendet, so dass beliebige Anwendungen ohne Anpassungen benutzt werden können. Als Problem erweist sich dabei, dass der Windows-Desktop stets monoskopisch ist, während das Stifterunternehmen 3DInsight GmbH vor allem an stereoskopischen Anlagen und Anwendungen interessiert ist. Deshalb entwickelten wir ein Verfahren, um eine monoskopische Textur als Kodierung stereoskopischer Formate zu interpretieren, und in geeignete andere Formate zu konvertieren. So unterstützen die meisten Projektoren z.B. das Side-By-Side Format, bei dem die Bilder für das linke und rechte Auge nebeneinander als ein monoskopisches Bild (mit halbierter Auflösung) übertragen werden. Während ein solches Format als Ausgabe-Format zu den Projektoren hin sehr gut geeignet ist, ist es als Eingabe-Format sehr ungünstig: Wenn eine Anwendung 3D-Inhalte in einem Fenster präsentieren möchte, müsste sie dafür nun zwei Fenster öffnen, die sich in den unterschiedlichen Hälften des Desktops befinden müssten, und deren Größe und Position stets synchron gehalten werden muss. Keine gebräuchliche Anwendung unterstützt derartige Stereo-Modi. Als sinnvolles Eingabe-Format sehen wir dagegen das „Checkerboard“-Format, bei dem beide Ansichten schachbrettartig ineinander verschachtelt werden. Dann bleibt der Bildinhalt in einem Fenster. Weiterhin ist dieses Format bereits als natives Format für einige DLP-Projektoren gebräuchlich, weshalb es bereits direkt von Grafiktreibern unterstützt wird. Verwendet eine Anwendung die allgemeine Stereoskopie- Funktionalität der Render-Schnittstelle (z.B. „Quad Buffering“ bei OpenGL), so kann über den Grafiktreiber eine Ausgabe im Checkerboard-Format eingestellt werden. Auf diese Weise konnten wir demonstrieren, dass selbst ohne expliziten Stereoskopie-Support stereoskopische Anwendungen ohne Anpassungen betrieben werden können.

4. Separate Verarbeitungshardware:
   Für die Umsetzung wurde auf Video-Capture Hardware zurückgegriffen, welche wie ein Monitor HDMI, DVI oder DisplayPort-Signale empfangen kann. Ein weiterer PC kann dann zwischen den Anwendungsrechner und die Projektionsanlage geschaltet werden, und die Kalibrierung durchführen. Die Projektionsanlage verhält sich dann wie ein „großer Monitor“. Es können beliebige Geräte als Signalquellen dienen, nicht nur PCs, und es gibt keine Einschränkungen hinsichtlich zu verwendender Betriebssysteme oder Render-Schnittstellen. Allerdings genügt zur Ausnutzung der vollen Auflösung häufig nicht ein einziges Signal, sondern es müssen mehrere Signale kombiniert werden. Dies führt allerdings zu hohen Anforderungen hinsichtlich der zu verarbeitenden Bandbreite.

Die verschiedene Ansätze unterscheiden sich stark hinsichtlich Kriterien wie Aufwand und Praktikabilität (z.B. Notwendigkeit der Anpassung jeder einzelnen Applikation auf Ebene 1), Performanceverhalten (wie z.B. Darstellungslatenzen) oder Nutzerfreundlichkeit. Keiner der Ansätze kann alle Kriterien gleichzeitig optimal erfüllen, so dass je nach Anwendungsszenario und Einsatzzweck eine geeignete Ebene gewählt werden muss. Im Rahmen des Projekts erfolgte eine Proof-of-Concept-Implementation aller vier Ebenen, um alle Ansätze umfassend beurteilen zu können.

Für den Arbeitspunkt Echtzeit-Videoverarbeitung multipler Signalquellen wurden zunächst die Anforderungen sowie die zu unterstützenden Anwendungsszenarien mit dem Stifterunternehmen 3DInsight abgestimmt. Als Referenz für die Gesamtauflösung dient eine Powerwall mit sechs FullHD Projektoren (5760 x 2160 @120Hz) im Stereobetrieb. Das Modul soll stereoskopische Signale verschiedener Formate, wie Displayport oder HDMI und eine Synchronisation mit unterschiedlichen Synchronisationsgruppen unterstützen. Die verschiedenen Signalquellen sind intern an ein Video Capture Board angeschlossen und müssen entsprechend aufbereitet und auf die Graphikkarte übertragen werden. Die Standardverarbeitung sieht vor, dass die Daten der eingehenden Signale vor der Übertragung auf die Graphikkarte im Arbeitsspeicher zwischengespeichert werden. Bei ersten Tests zeigte sich jedoch, dass die Übertragungsrate des internen Bussystems zu langsam ist, um die Datenmenge in Echtzeit zu verarbeiten. Beim Zuspielen von zwei Stereosignalen mit einer Auflösung von je 2880 x 2160 @60Hz und einem zusätzlichen FullHD Mono-Signal (1920 x 1080 @60Hz) für die Bild-in-Bild Darstellung müssen ca. 1,6 GigaPixel pro Sekunde ($\approx 6,4GB/s$) übertragen und verarbeitet werden. Eine effizientere Datenübertragung bietet der DMA-Modus des Busmasters, bei dem die Daten direkt vom Video Capture Board in den Graphikspeicher übertragen werden. Dies erfordert jedoch spezielle Hardware- und Treiberunterstützung. Der Graphiktreiber einer AMD FirePRO unterstützt durch den DirectGMA-Modus diese Art der Datenübertragung. Allerdings wird

dabei der zugreifbare Speicher auf 128MB beschränkt. Im genannten Anwendungsszenario wird für eine effiziente Verarbeitung und Synchronisation mindestens Tripple Buffering benötigt und somit zwischen 310MB und 520MB Graphikspeicher. Um diese Diskrepanz zu lösen, hat das Stifterunternehmen 3DInsight Kontakt mit AMD aufgenommen. Daraufhin hat AMD uns ein spezielles Video-Bios zur Verfügung gestellt, bei dem das Limit des zugreifbaren Speichers auf 768MB erhöht wurde. Der Nachteil dieser Custom-Lösung ist jedoch, dass die Treiber nur mit einer bestimmten Kombination von Mainboard und Graphikkarte funktionieren. Es wurde jedoch erreicht, dass künftige AMD Graphikkarten generell über mindestens 512MB DMA-Speicher verfügen werden.

Bei der Verarbeitung der Eingangssignale müssen die verschiedenen Bild-Modi erkannt und verarbeitet werden. Signale im Mono- sowie dem Side-by-Side-Bildmodus können direkt in Bezug auf die Zuordnung der Framebuffer für das linke und rechte Auge verarbeitet werden. Problematischer ist die Verarbeitung von Frame-Sequential Stereo-Signalen, da hierbei abwechselnd die Bilder für das linke/rechte Auge übertragen werden. Als einzige der betrachteten Signalquellen kodiert der DisplayPort ein Flag für die Bildzuordnung im Signalstream. Dieses kann vom Video-Grabber ausgelesen und ausgewertet werden. Bei allen anderen Signalquellen ist dies nicht der Fall. Die Zuordnung der Bilder (links/rechts) ist somit nicht notwendigerweise bekannt. Probleme bereiten hier die Schätzung des initialen Zustands sowie die Nachverfolgung von Frame-Drops. Als Proof-of-Concept wurde dazu eine Software implementiert, die Hilfsinformationen direkt in den Randpixeln des Bildes kodiert. Das erfordert jedoch die Installation einer zusätzlichen Software auf dem Zuspiel-System.

2.6.5 Erfassung von Nutzerposition und -eingaben

In Zusammenarbeit mit dem Mitarbeiter AB1 wurde eine Testumgebung geschaffen, die die Erfassung und Verarbeitung von Nutereingaben ermöglicht. Der Nutzer wird dazu mit Hilfe einer Webcam erfasst. Die Menscherkennung erfolgt mit OpenPose. Bei OpenPose handelt es sich um ein real-time object detection system auf Basis eines Convolutional Networks. Als Ergebnis liefert die Bibliothek ein vereinfachtes menschliches Skelett in Form einer Knochenhierarchie.

Auf Basis der räumlichen Position der aufzeichnenden Kamera sowie des relativen Winkels zwischen Blickrichtung der Kamera und der Ebene des Fußbodens erfolgt die Bestimmung der Nutzerposition durch räumliche Triangulierung.

Für die Demonstration der Nutzerinteraktion wurde ein Versuchsaufbau entwickelt, in dem ein Nutzer Schach gegen den Computer spielen kann. Die Interaktion erfolgt mittels sta-

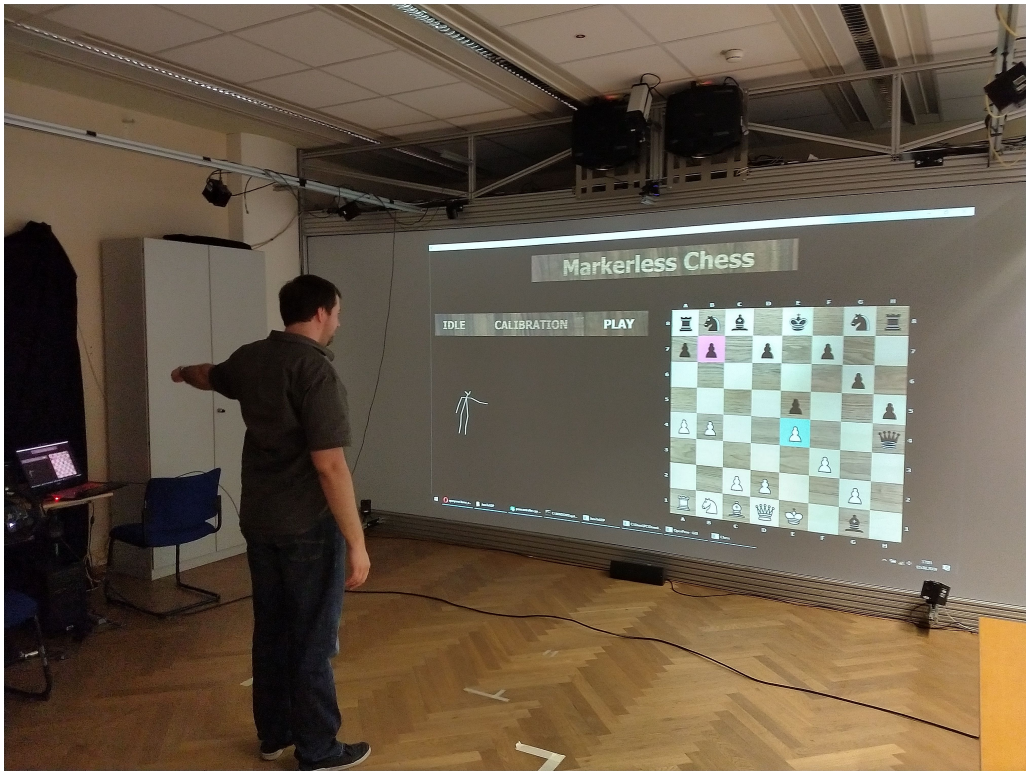


Abbildung 82: Darstellung der entwickelten Beispielanwendung zur Demonstration von Nutzerinteraktionen.

tischer Posen. Die Züge der Schachfigur werden durch Veränderung der Nutzerposition gesteuert. Abbildung 82 zeigt einen Überblick der Testumgebung.

2.6.6 Bilanz nach Erreichen des letzten Meilensteins in AB5

Mit dem Erreichen des letzten Meilensteins wurden die entwickelten Verfahren als Funktionsmuster implementiert und dem Stifterunternehmen 3DInsight zur Verfügung gestellt. Zusammen mit dem Stifterunternehmen 3DInsight GmbH wurden die erreichten Ergebnisse begutachtet und ausgewertet.

2.7 Notwendigkeit und Angemessenheit der geleisteten Arbeit (3.)

Die erzielten Ergebnisse in den fünf inhaltlichen Arbeitsbereichen (siehe Abbildung 1) konnten nur durch die gezielte Förderung einer personell stark aufgestellten Nachwuchsforschergruppe erzielt werden. Die Bearbeitung der Forschungsfragen war durch die KMU aufgrund des Risikos und der notwendigen Vorarbeiten nicht zu stemmen. Zudem waren die anvisierten technologischen und wissenschaftlichen Fortschritte in allen Einzelbereichen enorm hoch gesteckt. Diese Kombination der Untersuchung komplexer und heterogener

Forschungsfragen, die die gesamte Wertschöpfungskette der Medienproduktion berücksichtigen, wäre ohne die gezielte Förderung der Initiative LocalizeIT so weder für die KMU noch für die TU Chemnitz möglich gewesen.

Die im Projektverlauf entwickelten Lösungsansätze erforderten intensive Betrachtungen grundlegender Fragen in den Bereichen Retrieval, Signalverarbeitung (in Bild und Ton), maschinelles Lernen, insbesondere *Deep Learning*, die ohne die Förderung von LocalizeIT in diesem Maße nicht realisierbar gewesen wäre. Die entwickelten Funktionsmuster zur automatischen Klassifikation von Objekten und Personen und die dazu entwickelte technische Infrastruktur (siehe Abschnitt 2.2.7) ermöglichten es den Partnerunternehmen ihre bestehenden Geschäftsfelder an zukünftige Entwicklungen der Methoden künstlicher Intelligenz (KI) bzw. moderner maschineller Lernverfahren heranzuführen. Durch den anwendungsdomänenübergreifenden Ansatz in LocalizeIT konnten so Fragestellungen der Leistungsfähigkeit und Generalisierbarkeit erörtert werden, die für die Beteiligten KMU von wesentlicher Bedeutung waren vor dem Hintergrund dass KI gerade in aller Munde ist. Neben den inhaltlichen Arbeiten haben die Nachwuchswissenschaftler aktiv an der Entwicklung und Umsetzung von Maßnahmen zur Erreichung der förderpolitischen Ziele des Programms InnoProfile gearbeitet. Diese vielfältigen Aufgaben waren zwar implizit bereits im Antrag vorgesehen, resultierten in Summe aber letztlich in einem deutlichen Mehraufwand. Ferner wurden zur Durchführung dieser Maßnahmen keine Mittel bereitgestellt. Sowohl die vorgenommenen Investitionen als auch die erzielten Forschungsergebnisse wurden in die Infrastruktur am Standort in Chemnitz integriert. Die gewonnenen Erkenntnisse wurden in ihren wesentlichen Teilen in die Lehre der Stiftungsjuniorprofessur eingearbeitet, was im wesentlichen in der Vielzahl studentischer Arbeiten widerspiegelt. Damit wird die Professur Stiftungsjuniorprofessur Media Computing bzw. die TU Chemnitz als kompetenter Ansprechpartner für privatwirtschaftliche Unternehmen etabliert und nachhaltig gestärkt, für die Maschinelles Lernen, Audio- und Videoverarbeitung Teil ihrer Zukunftsstrategie ist. Die eigens durchgeführten nationalen Workshops, die Beiträge zu internationalen Konferenzen sowie die hervorragenden Ergebnisse bei internationalen Vergleichskampagnen im Bereich Audio- und Video-Klassifikation haben zur Stärkung und Schärfung dieses Profils beigetragen.

2.8 Nutzen/Verwertbarkeit der Ergebnisse (4.)

Ziel des Vorhabens LocalizeIT war es Frameworks und Algorithmen für konkrete Anwendungsszenarien zu entwickeln, die bspw. von den Partnerunternehmen zur Produktreife geführt werden können. Demnach sind folgende wesentliche Algorithmen entstanden, deren Anwendungspotenzial im Rahmen exemplarischer Prototypen für konkrete Einsatzszenarien aufgezeigt werden konnte. Im Folgenden werden einige Ergebnisse und deren

Verwertungsstrategie und -möglichkeiten vorgestellt, Potenziale und Konzepte für Aus- oder Neugründungen aufgezeigt und FuE-bezogene Abstimmungen innerhalb der Hochschule(n), Fraunhofer ENAS und mit umliegenden KMU bzw. der Industrie dargestellt:

- Audioklassifikator für Vogelgesang - Die Zweitplatzierung im BirdCLEF-Wettbewerb 2017 und die Ergebnisse in der Dissertation von Stefan Kahl sind Ausdruck, dass im Projekt eine Nischendomäne identifiziert wurde auf der wir Spitzenergebnisse erzielen konnten. Über diese Vorarbeiten konnte bereits eine Postdoc-Stelle für Stefan Kahl über Stifter des Cornell Lab of Ornithology für 2 Jahre in Deutschland finanziert werden. Das Interesse die BirdNET-App über den Android-Store hinaus weiterzuentwickeln wurde deutlich durch die Anfragen von Apple, aber auch von Umweltschutzorganisationen und Konzernen wie Nestlé, die Interesse an einem grünen Image/Umweltmarketing haben. Die Verwertbarkeit wird aktuell von Stefan Kahl vorangetrieben durch die Weiterentwicklung der Vogelgesangserkennungsalgorithmik, aber auch der Anwendungen z.B. für IPhone und die o.g. Organisationen.

- Audioklassifikator für Ambient-Assisted-Living Kontext - Die Vorarbeiten haben gezeigt, dass für den Wohnkontext relevante Audioklassen zunehmend genau identifiziert werden können und auch portierbar sind auf mobile Geräte. Vor dem Hintergrund aktueller Ausgründungsprojekte an der TU Chemnitz im Bereich Smart Sensor (Verhaltensanalyse und Gaserkennung, siehe Kapitel 2.9) ist es denkbar, Audioanalysen in solche Sensoren zu integrieren. Auf Interesse stießen diese Ergebnisse aber auch in der Professur Schweißtechnik und deren industriellen Partner, die auf der Suche nach qualitätssichernden Maßnahmen in Schweißprozessen auch Mikrofone als Sensoren erwägen. Es wird mittelfristig angestrebt die Forschung auf diesem Gebiet gemeinsam voranzutreiben, z.B. durch gemeinsame ZIM-Projekte (Zentrales Innovationsprogramm Mittelstand). Die Corant GmbH, eine Ausgründung der TU Chemnitz, entwickelt Gassensoren (air-Q) zur Kontrolle der Luftqualität von Innenräumen. Interesse wurde bekundet an KI-basierten Verfahren. Auch die spätere Erweiterung des ohnehin modularen Sensors auf Audiogeräuscherkennung ist denkbar. Gründer- und Forschungsförderung über das SMWK (Sächsisches Ministerium für Wissenschaft und Kunst) wäre denkbar.

- Klassifikationsframework für Massendaten und *Instance Search* - Die Architektur des für den TRECVID-Wettbewerb entwickelten Systems hat eindrücklich gezeigt, wie Webtechnologien eingesetzt werden kann um einerseits Prozesse zu verteilen, andererseits aber auch entsprechende Maschinelle Lernframeworks einzubinden und das ganze auf Datenbankebene massendatentauglich zu machen. Die demonstrierten Ergebnisse haben Interesse geweckt in der Medizin, konkret der Augenmedizin. Dort

gibt es noch viele ungeklärte Forschungsfragen zur Wirksamkeit von Medikamenten, Therapien und zu Fragen der individuellen Medizin. So hat die Novartis GmbH bereits ein Anschubprojekt finanziert mit Hilfe dessen in medizinischen Massendaten nach Ursachen zur Medikamentenwirksamkeit geforscht werden soll. Mit TOPOs[22] konnte gemeinsam mit Prof. Ritter ein BMBF-Verbund-Projekt dazu eingeworben werden mit einem Umfang von über 2.000.000€. Auch die SAB hat ein Projekt im Augenmedizin-Forschungsfeld bereits gefördert und die Juniorprofessur Media Computing ist mit einem Trail zum Thema Auge 4.0 Teil des Smart Systems Hub und steht für Kompetenzen in der Massendatenverarbeitung und Maschinellen Lernverfahren. Das Forschungsthema wird zukünftig weiterentwickelt, da es sowohl Interesse auf Seiten der Pharmaindustrie als auch der Medizingerätehersteller, aber auch beim Freistaat Sachsen (Verbesserung der Patientenversorgung im ländlichen Raum) gibt. Gemeinsam mit dem Klinikum Chemnitz wird aktuell versucht ein InnoExpert zu beantragen zur Etablierung eines Klinischen Ökosystems für die Region Chemnitz (Ziel: SMWK-Förderung). Über dieses Netzwerk ist aktuell ein BMBF-Antrag gemeinsam mit Fraunhofer ENAS und dem Klinikum Chemnitz auf einem weiteren klinischen Forschungsthema in Vorbereitung.

- Klassifikator für Chip-Fehlererkennung - Nach unseren Erkenntnissen haben wir den weltweit besten Klassifikator für heterogene Halbleiter-Wafer-Defekte. Wenngleich die Erkennungsraten noch nicht industrietauglich sind, ist das Potenzial hoch, dass sie es künftig werden, wenn es mehr Trainingsdaten gibt oder in homogeneren Kontexten weniger diverses Bildmaterial entsteht. Der Stifter 3D-Micromac hat angedeutet mittelfristig Kontakt zu interessierten Kunden herzustellen, die CNN-basierte Klassifikation zur Qualitätskontrolle einsetzen werden. Ferner zeigte sich die Firma offen weiter zu Geld zu stiften z.B. im Rahmen einer Promotion nach dem Vorbild der Landesinnovationspromotionen.

- Klassifikator für Fahrraderkennung - Es zeigt sich, dass moderne Objekterkennungsframeworks anpassbar sind, durch *Transfer Learning* und durch Re-Training mit zielgerichteten Daten. So können z.B. Schwächen aufgrund der Perspektive effektiv kompensiert werden. Damit können Überwachungssensoren um smarte Funktionen erweitert werden, indem sie konkrete Objekte erkennen, zählen und ggf. Meldungen an ein externes System weitergeben. Diese Erkenntnisse könnten bei der Hardwarekonzeption neue Versionen von bereits smarten Sensor-Systemen wie dem S2000 der Intenta GmbH berücksichtigt werden und damit neue Märkte erschließen, z.B. die

[22] `https://topos.averbis.de/`, Therapievorhersage durch Analyse von Patientendaten in der Ophthalmologie

Innenraumüberwachung öffentlicher Verkehrsmittel aus Sicherheitsaspekten heraus oder zur verbesserten Assistenz für Kunden.

- Klassifikator für Personenzählung - Die Arbeiten von Falk Schmidsberger et al. zeigte wie Sensor- und Algorithmen-Fusion dazu beitragen können Aufgaben wir Personenzählung zuverlässiger zu übernehmen. Eine Weiterentwicklung im Rahmen vernetzter Sensoren ist denkbar. Die Umsetzbarkeit solcher Technologien und deren Algorithmen wird in mittlerer Zukunft weiter diskutiert. Die Intenta GmbH hat sich offen gezeigt auch weiterhin Daten für die Unterstützung von Promotionen auf dem Gebiet zur Verfügung zu stellen.

Die Partnerfirmen Intenta GmbH und 3D-Micromac AG unterstützten bereits zum aktuellen Zeitpunkt Drittmittelanträge der Stiftungsjuniorprofessur Media Computing, mit denen zuerst spezielle Nachwuchsqualifizierungsprofile und -kompetenzen hervorgehen und damit eine Infrastruktur für potentielle Innovationen entstehen kann. Zu den unterstützten und bewilligten Anträgen zählt die ESF-Landesinnovationspromotion für Tobias Schlosser zum Thema „Robuste Objektklassifikation zur klinischen sowie industriellen Bild- und Videoverarbeitung auf Basis der hexagonalen Bildrasterung für maschinelle Lernverfahren“, welche die Ausrichtung des Forschungsprofils der Juniorprofessur in Richtung Maschinelles Lernen und Künstlicher Intelligenz forcieren soll und die Stifterforschungsthemen über die Projektlaufzeit hinaus vertreten wird. Im Fokus stehen dabei die Lokalisierung von Pathologien in der Ophthalmologie, die Klassifikation in der Laserverarbeitung und Objekterkennung im Automotive und AAL-Bereich (*Ambient Assisted Living*).

Die im Projekt beschaffte Infrastruktur, insbesondere der Rechen-Cluster mit 190TB Storage, die GPU-Workstations und das Labor werden weiterhin genutzt durch die Juniorprofessur Media Computing und stellen eine notwendige und leistungsstarke Infrastruktur bereit, die es erlaubt sich in neuen FuE-Vorhaben zu beteiligen, insbesondere dann wenn es um Massendaten und Entwicklung von Algorithmen auf Basis maschineller Lernmethoden handelt.

Die Neuausschreibung und Weiterführung der Stiftungsjuniorprofessur Media Computing (bei Übernahme der Kosten durch die Fakultät für Informatik) für weitere 4-6 Jahre über die Projektlaufzeit hinaus wäre ohne die Projektförderung in LocalizeIT nicht möglich gewesen und zeigt, dass die Profilbildung und erzielten Erfolge von der Fakultät für Informatik und damit auch von der TU Chemnitz unterstützt werden.

Der Forschungsbereich Biochemie und Analyse von RNS-Interaktion mit Prof. Dr. Roland Sigel (Universität Zürich, Institut für Chemie, Metallo-RNA-Arbeitsgruppe) führte zu zahlreichen gemeinsamen Publikationen [31, 52, 30, 116]. Aktuell werden Kooperationsmöglichkeiten sondiert um gemeinsam eine Evaluationskampagne durchzuführen im Be-

reich molekulare Sortierung in Klassen. Konkrete Anträge zur Weiterentwicklung des Forschungsgebietes über DFG sind mittelfristig geplant.

Unter den Publikationen von Robert Manthey und Tobias Schlosser finden sich einige zum Thema Hexagonale Bildverarbeitung. Auch hier ist eine weitere Vertiefung dieses eher grundlegenden Forschungsthema aus DFG-Projektebene denkbar.

Beitrag des Vorhabens zur Herausbildung eines besonderen Forschungs- und Innovationsprofils in der betreffenden Region und Überregional

Durch die Förderung im Projekt LocalizeIT konnte ein vertieftes Profil in Methoden der Künstlichen Intelligenz, insbesondere dem überwachten Lernen mit neuronalen Faltungsnetzen entwickelt werden, da diese in allen 5 Arbeitsbereichen inhaltliche Anteile hatten. Die erworbenen Kompetenzen der mehr als 10 Projektmitarbeiter inkl. den Projektleitern führte zu weitreichenden Kontakten von Chemnitz über Mittweida in die Welt. Kontakte und Vernetzung gab es mit Gruppen der Universität Zürich und der Cornell University, mit diversen Augenkliniken in Deutschland (Chemnitz, Zschopau, Freiburg, Greifswald), aber auch zu lokalen Firmen (IMM, Fusion Systems, Baselabs, IAV, Thalheim Spezialoptik, Biostep GmbH, Geos, ...) und großen Unternehmen (Novartis, Bayer, Zeiss, Heidelberg Engineering, Nestlé, Apple, Google). Über Prof. Marc Ritters Berufung an die HS Mittweida und die Weiterentwicklung der KI-Vertiefung ist die Region Chemnitz/Mittweida nun deutlich sichtbarer und kann Wissens- und Technologietransfer sowohl lokal, aber auch international zukünftig noch besser leisten. Über die genannten Kontakte und die gemeinsam organisierten Promotionsworkshops von Medieninformatikern der TU Chemnitz, HS Mittweida und des Medienzentrums der TU Dresden hat die Juniorprofessur das Profil der Informatik auf mehr Interdisziplinarität ausgerichtet. Insbesondere im Bereich Lebenswissenschaften und digitale Medizin ist die Juniorprofessur Media Computing dank der Vorarbeiten während und nach LocalizeIT ein regionaler Ansprechpartner und Netzwerker für FuE-Vorhaben und Technologieentwicklung.

Die Juniorprofessur Media Computing wurde in den Forschungsschwerpunkt „Intelligente Multimediale Systeme“ der Fakultät für Informatik aufgenommen. Dort ergänzt sie insbesondere die Professur Künstliche Intelligenz durch stärker anwendungsbezogene Forschung und Algorithmenentwicklung im Bereich überwachten Lernens. Ferner öffnet sie den Forschungsschwerpunkt über die Grenzen der Informatik hinaus, insbesondere um neue Anknüpfungspunkte in den Lebenswissenschaften zu finden, z.B. der Augenmedizin und der Bioinformatik im Bereich RNS-Faltung (RNS - Ribonukleinsäure).

Strategie zur Personalqualifizierung und Nachwuchsgewinnung für die umliegenden Unternehmen

Im Berichtszeitraum wurden spezifische Weiterbildungsaktivitäten für Unternehmen durch regelmäßige gemeinsame Progress-Meetings mit dem wissenschaftlichen Personal der IPT-Initiative angeboten. Das beinhaltete auch Übersichtsvorträge zu ausgewählten Themen, die fachspezifischen State of the Art behandeln, z.B. *Deep Learning*. Weiterhin wurden eine Reihe von Masterarbeiten mit Firmen betreut und durchgeführt. Zunehmend kommen auch Anfragen und Möglichkeiten für Weiterbildungsangebote an Ärzte (z.B. Ophthalmologisches Herbstmeeting).

Wie dem Kapitel 2.10.3 zu sehen wurden über über 90 studentische studentische Arbeiten betreut, davon waren gut 20% gemeinsam betreute Arbeiten mit Unternehmen und weniger als 5 in Zusammenarbeit mit einem Fraunhofer Institut. Nahezu alle Absolventen berichteten davon bereits eine Arbeitsstelle in Aussicht zu haben. Dazu gehörten häufig Firmen aus dem Automotive-Bereich aber auch Web-Engineering. Mehr als 10 Absolventen konnten auch als Doktoranden bzw. für eine akademische Laufbahn gewonnen werden.

Neben den im Antrag unter Kapitel 2.10.2 erwähnten Doktoranden konnten weitere 8 Dissertationsvorhaben mit den Vorarbeiten des Projekts durch Prof. Ritter an der HS Mittweida eingeworben und begonnen werden und werden Teils in Kooperation mit Prof. Eibl von der Professur Medieninformatik der TU Chemnitz durchgeführt. An den Namen ist zu erkennen, dass es viele Koautoren von Publikationen sind, die im Rahmen von LocalizeIT entstanden:

1. R. Vogel, “tba,” PhD Thesis, Hochschule Mittweida, Chemnitz.
2. R. Thomanek, “tba,” PhD Thesis, Hochschule Mittweida, Mittweida.
3. C. Roschke, “tba,” PhD Thesis, Hochschule Mittweida, Mittweida.
4. T. Rolletschke, “tba,” PhD Thesis, Hochschule Mittweida, Mittweida.
5. M. Hill, “tba,” PhD Thesis, TU Chemnitz, Chemnitz.
6. M. Heinzig, “tba,” PhD Thesis, TU Chemnitz, Mittweida.
7. R. Hasan, “tba,” PhD Thesis, Hochschule Mittweida, Mittweida.
8. M. Benndorf, “tba,” PhD Thesis, TU Chemnitz, Chemnitz.
9. C. Hösel, “tba,” PhD Thesis, TU Chemnitz, Chemnitz.

Für die beiden Projektleiter Marc Ritter (08/2014-09/2016) und Danny Kowerko (09/2016-07/2019) diente LocalizeIT der Weiterqualifizierung in ihrer akademischen Laufbahn. Beide erreichten dadurch für sich den nächsten Schritt, Marc Ritter durch die Berufung auf eine W2-Professur Medieninformatik an der HS Mittweida und Danny

Kowerko durch die Berufung auf die W1-Juniorprofessur Media Computing an der TU Chemnitz.

2.9 Bekanntwerden relevanter Ergebnisse Dritter (5.)

Im Berichtszeitraum sind keine relevanten Ergebnisse Dritter bekannt geworden.

- 3D Visionslab GmbH, Ausgründung im Smart Sensor Bereich an der TU Chemnitz - Die 3D Visionslab GmbH entwickelt einen omnidirektionalen Stereo-Sensor mit 3 statt 2 Kamerasensoren zur Verhaltensanalyse. Die Hardware setzt bereits auf Nvidia-Technologie, so dass die Implementierung CNN-basierter Algorithmen möglich ist.
- Google/Deep Mind: Google bzw. die aufgekaufte Firma Deep Mind macht immer wieder Schlagzeilen mit neuen Durchbrüchen im Bereich künstlicher Intelligenz, sei es AlphaGo oder die entwicklung medizinischer Expertensysteme für die Augenmedizin ([37]).
- Digitalkonzerne Amazon, Google, Facebook, Microsoft - Einige der im Projekt benutzten Objekterkennungsframeworks wurden von großen Digitalkonzernen mit entwickelt. Zunehmend mehr von ihnen bieten Cloud-Lösunen an im Bereich *Deep Learning*, was auch Objekterkennungsframeworks einbezieht, die das angepasste Entwickeln neue Algorithmen für spezifische Klassifikationsaufgaben deutlich erleichtert.
- TRECVID-Teilnehmerfeld - Die Teilnehmer des TRECVID-Wettbewerbs tw. aus dem großindustriellen Umfeld (Hitachi, IBM) sind immernoch deutlich überlegen wenn es um die Klassifikation von Personen und Objekten in großen Datenmengen geht.
- Fraunhofer Oldenburg - Im Raum Oldenburg bilden Hochschulen und Forschungsreinrichtungen einen Schwerpunkt schon in der Ausbildung aber auch in der Forschung rund um das Thema Audio. Das dort ansässige Fraunhofer IDMT hat ausgewiesene Kompetenzen in der audiobasierten Ortung mit Mikrofon-Arrays aber auch in der dazugehörigen Implementierung auf mobiler Rechentechnik.
- Museum für Naturkunde Berlin - Mario Lasseck ist der ausgewiesene Experte für Vogelstimmenklassifikation. Er hat mehrfach den BirdCLEF-Wettbewerb gewonnen.

2.10 Veröffentlichungen und Publikationen (6.)

Projektrelente Publikationen sind auf einer Zotero-Webseite[23] digital zugänglich. Folgende Publikationen wurden unter Mitwirken von Mitarbeitern der IPT-Initiative veröffentlicht:

2.10.1 Publikationen

1. F. D. Steffen, M. Khier, D. Kowerko, R. A. Cunha, R. Börner, and R. K. O. Sigel, "Metal ions and sugar puckering balance single-molecule kinetic heterogeneity in RNA and DNA tertiary contacts," Nat Commun, vol. 11, no. 1, p. 104, Dec. 2020, doi: 10.1038/s41467-019-13683-4.

2. F. Schmidsberger and D. Kowerko, "Objektverfolgung durch Sensorfusion in Multisensorumgebungen," in Chemnitzer Informatik Berichte 2020, Chemnitz, 2020, vol. CSR-20-01, pp. 82–96.

3. T. Schlosser, F. Beuth, M. Friedrich, and D. Kowerko, "Fehlerdetektion und -klassifikation bei Laserschneidprozessen mittels Deep Neural Networks," in Chemnitzer Informatik Berichte 2020, Chemnitz, 2020, vol. CSR-20-01, pp. 67–81.

4. A. Sampath-Kumar, R. Erler, and D. Kowerko, "CNN-based Audio Classification for Environmental Sounds, Ambient Assisted Living and Public Transport Environments using an Extensive Combined Datas," in Chemnitzer Informatik Berichte 2020, Chemnitz, 2020, vol. CSR-20-01, pp. 29–66.

5. T. Kretzschmar and D. Kowerko, "Image related metadata generation, storage and retrieval for big datasets," in Chemnitzer Informatik Berichte 2020, Chemnitz, 2020, vol. CSR-20-01, pp. 9–28.

6. V. Forch, J. Vitay, and F. Hamker, "Recurrent Spatial Attention for Facial Emotion Recognition," in Chemnitzer Informatik Berichte 2020, Chemnitz, 2020, vol. CSR-20-01, pp. 1–9.

7. R. Thomanek et al., "University of Applied Sciences Mittweida and Chemnitz University of Technology at TRECVID Instance Search 2019," in TRECvid Workshop Proceedings 2019, Gaithersburg, UNITED STATES, 2019, p. 9.

8. R. Thomanek et al., "University of Applied Sciences Mittweida and Chemnitz University of Technology at TRECVID ActEv 2019," in TRECvid Workshop Proceedings 2019, Gaithersburg, UNITED STATES, 2019, p. 7.

[23] https://www.zotero.org/groups/1046750/localizeit/

9. R. Thomanek et al., “A Scalable System Architecture for Activity Detection with Simple Heuristics,” in 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 2019, pp. 27–34, doi: 10.1109/WACVW.2019.00012.

10. S. Taubert, S. Kahl, D. Kowerko, and M. Eibl, “Automated Lifelog Moment Retrieval based on Image Segmentation and Similarity Scores,” in Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, 2019.

11. T. Schlosser, F. Beuth, M. Friedrich, and D. Kowerko, “A Novel Visual Fault Detection and Classification System for Semiconductor Manufacturing Using Stacked Hybrid Convolutional Neural Networks,” in 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Zaragoza, Spain, 2019, pp. 1511–1514, doi: 10.1109/ETFA.2019.8869311.

12. A. Sampath Kumar, R. Erler, and D. Kowerko, “A Real-Time Demo for Acoustic Event Classification in Ambient Assisted Living Contexts,” in Proceedings of the 27th ACM International Conference on Multimedia - MM ’19, Nice, France, 2019, pp. 2205–2207, doi: 10.1145/3343031.3350600.

13. R. Manthey et al., “Visual System Examination using Synthetic Scenarios,” in Proc. IHSI 2019, San Diego, CA, 2019, pp. 1–5, doi: 10.1007/978-3-030-11051-2_63.

14. R. Manthey et al., “An Exploratory Inspection of the Detection Quality of Pose and Object Detection Systems by Synthetic Data,” in HCI International 2019 - Posters, Cham, 2019, pp. 287–294, doi: 10.1007/978-3-030-23528-4_40 ER.

15. R. Manthey and D. Kowerko, “Hexagonal Image Generation by Virtual Multi-Grid-Camera,” in Advances in Intelligent Systems and Computing, San Diego, CA, 2019, vol. 903, pp. 1–8, doi: 10.1007/978-3-030-11051-2_3.

16. R. Manthey, “Bilderzeugung von virtuellen Szenen mittels vier- und sechseckbasierter oder wahlfreier Rasterung,” tm - Technisches Messen, vol. 0, no. 0, Jun. 2019, doi: 10.1515/teme-2019-0027.

17. R. Thomanek et al., “University of Applied Sciences Mittweida and Chemnitz University of Technology at TRECVID 2018,” in TRECvid Workshop Proceedings 2018, Gaithersburg, UNITED STATES, 2018, pp. 1–17.

18. S. Taubert, M. Mauermann, S. Kahl, D. Kowerko, and M. Eibl, “Species Prediction based on Environmental Variables using Machine Learning Techniques,” in Working notes of CLEF, 2018.

19. R. Manthey, R. Thomanek, C. Roschke, M. Ritter, and D. Kowerko, “Synthetic Ground Truth Generation for Testing, Technology Evaluation and Verification (SyntTEV),” in Proceedings of British HCI 2018, Belfast, UK, 2018.

20. R. Manthey and D. Kowerko, “Visuelle Szenendiskretisierung mittels wahlfreien Bildrastern am Beispiel von Drei-, Vier- und Sechseckpixeln,” in Forum Bildverarbeitung 2018, Karlsruhe, Germany, 2018, pp. 95–104.

21. S. Kahl, T. Wilhelm-Stein, H. Klinck, D. Kowerko, and M. Eibl, “Recognizing Birds from Sound - The 2018 BirdCLEF Baseline System,” CoRR, vol. abs/1804.07177, 2018.

22. S. Kahl, T. Wilhelm-Stein, H. Klinck, D. Kowerko, and M. Eibl, “BirdNET: Real-time Bird Sound Identification using Convolutional Neural Networks,” in Proceedings of the 10th International Conference on Ecological Informatics, Jena, 2018.

23. S. Kahl, T. Wilhelm-Stein, H. Klinck, D. Kowerko, and M. Eibl, “A Baseline for Large-Scale Bird Species Identification in Field Recordings,” in Working notes of CLEF, 2018.

24. M. Heinz, D. Kowerko, and G. Brunnett, “Camera-based color measurement of DLP projectors using a semi-synchronized projector camera system,” 2018, p. 23, doi: 10.1117/12.2307119.

25. J. Haupt, S. Kahl, D. Kowerko, and M. Eibl, “Large-Scale Plant Classification using Deep Convolutional Neural Networks,” in Working notes of CLEF, 2018.

26. M. C. A. S. Hadzic, R. Börner, S. L. B. König, D. Kowerko, and R. K. O. Sigel, “Reliable State Identification and State Transition Detection in Fluorescence Intensity-Based Single-Molecule Förster Resonance Energy-Transfer Data,” The Journal of Physical Chemistry B, vol. 122, no. 23, pp. 6134–6147, Jun. 2018, doi: 10.1021/acs.jpcb.7b12483.

27. H. Goeau, S. Kahl, H. Glotin, R. Planque, and A. Joly, “Overview of BirdCLEF 2018: monospecies vs. soundscape bird identification,” in CLEF 2018 Working Notes, 2018, p. 12.

28. R. Börner, D. Kowerko, M. C. A. S. Hadzic, S. L. B. König, M. Ritter, and R. K. O. Sigel, “Simulations of camera-based single-molecule fluorescence experiments,” PLOS ONE, vol. 13, no. 4, pp. 1–23, Apr. 2018, doi: 10.1371/journal.pone.0195277.

29. M. Vodel and M. Ritter, “Thermale Fingerabdrücke für Software Tasks,” presented at the Innosecure 2015 - Innovationen in den Sicherheitstechnologien, Deutschland, 2017.

30. M. Vodel and M. Ritter, “Thermal Fingerprints for Computational Tasks – Benefits and Security Issues,” in International Conference on Electronics, Information and Communication ICEIC, Thailand, 2017.

31. B. Meyer-Sickendiek, H. Hussein, and T. Baumann, “Rhythmicalizer,” in INFORMATIK 2017, Chemnitz, 2017, pp. 2189–2200, doi: 10.18420/in2017 218.

32. R. Manthey, T. Schlosser, and D. Kowerko, “Generation of Images with Hexagonal Tessellation using Common Digital Cameras,” in IBS Scientific Workshop Proceedings, 2017, vol. 4, pp. 47–50.

33. D. Kowerko, M. Rößner, S. Kahl, R. Herms, M. Eibl, and K. Engelmann, “Aufbereitung augenmedizinischer Bild-, Patienten- und Diagnosedaten zum Zwecke der Forschung - Ethikrichtlinien und deren praktische Umsetzung,” in Mensch und Computer 2017 - Workshopband, Regensburg, 2017, pp. 311–318, doi: 10.18420/muc2017-ws07-0288.

34. D. Kowerko, D. Richter, M. Heinzig, S. Kahl, S. Helmert, and G. Brunnett, “Evaluation of CNN-based algorithms for human pose analysis of persons in red carpet scenarios,” in INFORMATIK 2017, Chemnitz, 2017, pp. 2201–2209, doi: $10.18420/in2017_219$.

35. D. Kowerko, R. Manthey, M. Heinz, T. Kronfeld, and G. Brunnett, “Fast and accurate creation of annotated head pose image test beds as prerequisite for training neural networks,” in INFORMATIK 2017, Chemnitz, 2017, pp. 2221–2229, doi: $10.18420/in2017_221$.

36. D. Kowerko and S. Kahl, “WS34 - Deep Learning in heterogenen Datenbeständen,” in INFORMATIK 2017, Chemnitz, 2017, p. 2141, doi: $10.18420/in2017_213$.

37. T. Keller and D. Kowerko, “A web-based application for data visualisation and non-linear regression analysis including error calculation for laboratory classes in natural and life sciences,” in IBS Scientific Workshop Proceedings, Germany, 2017, vol. 4, pp. 12–15.

38. S. Kahl et al., “Large-Scale Bird Sound Classification using Convolutional Neural Networks,” in CEUR Workshop Proceedings (Working Notes of CLEF 2017 - Conference and Labs of the Evaluation), 2017, vol. 1866.

39. S. Kahl et al., “Technische Universität Chemnitz and Hochschule Mittweida at TRECVID Instance Search 2017,” TRECVid Workshop Proceedings, vol. 2017, pp. 1–7, Sep. 2017.

40. S. Kahl et al., “Acoustic Event Classification Using Convolutional Neural Networks,” in INFORMATIK 2017, Chemnitz, 2017, pp. 2177–2188, doi: $10.18420/in2017_217$.

41. B. John et al., “Quantification of geometric properties of the melting zone in laser-assisted welding,” in Proceedings of Lasers in Manufacturing 2017, München, 2017, pp. 1–9.

42. H. Hussein et al., “Design of a Laboratory for Audio and Video Based Object Localization and Tracking,” in IBS Scientific Workshop Proceedings, Laubusch, 2017, pp. 28–29.

43. M. Vodel and M. Ritter, “The TUCool Project - Low-cost, Energy-efficient Cooling for Conventional Data Centres,” presented at the 6th International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies (ENERGY 2016), Lisbon, 2016.

44. Timon Zietlow, Marcel Heinz, and Guido Brunnett, “Advanced luminance control and black offset correction for multi-projector display systems,” Journal of Virtual Reality and Broadcasting, vol. 12(2015), no. 4, Mar. 2016, doi: 10.20385/1860-2037/12.2015.4.

45. T. Schlosser, R. Manthey, and M. Ritter, “Entwurf und Implementierung von Optimierungs- und Funktionserweiterungen der hexagonalen Bildrasterung in der Videokompressionssoftware x264HMod,” in Studierendensymposium Informatik 2016 der TU Chemnitz, Chemnitz, 2016, pp. 75–85.

46. P. Rosenthal, M. Ritter, D. Kowerko, and C. Heine, “OphthalVis - Making Data Analytics of Optical Coherence Tomography Reproducible,” in EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3), Groningen, 2016, pp. 1–5, doi: 10.2312/eurorv3.20161109.

47. M. Ritter et al., “Simplifying Accessibility Without Data Loss: An Exploratory Study on Object Preserving Keyframe Culling,” in Universal access in human-computer interaction. users and context diversity, 2016, vol. 9739, p. 427408_1_En.

48. S. Müller, S. Kahl, and M. Eibl, “Automatisierte Qualitätsbeurteilung von (S)VHS-Digitalisaten Automated quality assessment of digitized (S)VHS-tapes,” in Forum Bildverarbeitung 2016, Karlsruhe, Germany, 2016, pp. 137–148, doi: 10.5445/KSP/1000059899.

49. D. Kowerko, M. Ritter, R. Manthey, B. John, and M. Grimm, “Quantifizierung der geometrischen Eigenschaften von Schmelzzonen bei Laserschweißprozessen,” in Forum Bildverarbeitung 2016, Karlsruhe, Germany, 2016, pp. 285–296, doi: 10.5445/KSP/1000059899.

50. T. Keller, D. Kowerko, and M. Ritter, “Entwicklung eines webbasierten Curvefitting Tools für komplexe Multiparameter-Funktionen,” in Studierendensym-

posium Informatik 2016 der TU Chemnitz, Chemnitz, 2016, pp. 75–85, doi: 10.13140/RG.2.1.3254.7448.

51. S. Kahl et al., “Technische Universitat Chemnitz at TRECVID Instance Search 2016,” TRECVid Workshop Proceedings, vol. 2016, pp. 1–8, Sep. 2016.

52. M. C. A. S. Hadzic, D. Kowerko, R. Börner, S. Zelger-Paulus, and R. K. O. Sigel, “Detailed analysis of complex single molecule FRET data with the software MASH,” in Proc. SPIE, 2016, p. 971119, doi: 10.1117/12.2211191.

53. R. Börner, D. Kowerko, H. G. Miserachs, M. F. Schaffer, and R. K. O. Sigel, “Metal ion induced heterogeneity in RNA folding studied by smFRET,” Coordination Chemistry Reviews, vol. 327–328, pp. 123–142, Nov. 2016, doi: 10.1016/j.ccr.2016.06.002.

54. G. Awad et al., “TRECVID 2016: Evaluating vdeo search, video event detection, localization, and hyperlinking,” in TRECvid Workshop Proceedings 2016, Gaithersburg, UNITED STATES, 2016.

55. S. L. Wood, G. S. Bahr, and M. Ritter, “Cognitive Tools for Design Engineers: A Framework for the Development of Intelligent CAD Systems,” i-com, vol. 14, no. 2, Jan. 2015, doi: 10.1515/icom-2015-0028.

56. M. Ritter et al., “Technische Universitat Chemnitz at TRECVID Instance Search 2015,” TRECVid Workshop Proceedings, Sep. 2015.

57. M. Ritter and G. S. Bahr, “An exploratory study to identify relevant cues for the deletion of faces for multimedia retrieval,” 2015, pp. 1–6, doi: 10.1109/ICMEW.2015.7169806.

58. M. Ritter, “Automated identification of blue and fin whale vocalizations using an ensemble-based classification system,” presented at the The 7th International DCLDE

59. Detection, Classification, Localization, and Density Estimation] Workshop, La Jolla, 2015.

60. R. Herms, D. Richter, M. Eibl, and M. Ritter, “Unsupervised Language Model Adaptation using Utterance-based Web Search for Clinical Speech Recognition.,” in CLEF (Working Notes), 2015.

61. M. Heinz and G. Brunnett, “Dense sampling of 3D color transfer functions using HDR photography,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, 2015, pp. 25–32, doi: 10.1109/CVPRW.2015.7301372.

62. M. Ritter et al., “Technische Universitat Chemnitz at TRECVID Instance Search 2014,” TRECVid Workshop Proceedings, Sep. 2014.

63. M. Ritter, “Towards a Sustainable Framework for the Analysis of Large Audiovisual Data Collections. In: Proceedings of Symposium on Computational Sustainabillity,” Görlitz, 2014, pp. 1–10.

64. S. Kahl, M. Ritter, and P. Rosenthal, “Automatisierte Beurteilung der Schädigungssituation bei Patienten mit altersbedingter Makuladegeneration (AMD),” in Proceedings of Forum Bildverarbeitung, Karlsruhe, 2014, pp. 179–190.

65. R. Herms, M. Ritter, T. Wilhelm-Stein, and M. Eibl, “Improving spoken document retrieval by unsupervised language model adaptation using utterance-based web search,” in INTERSPEECH-2014, Singapore, 2014, pp. 1430–1433.

Publikationen mit R. Sigel und R. Börner haben keinen direkten Projektbezug und sind auf die erwähnte Kooperation mit Prof. Sigels Labor an der Uni Zürich zurückzuführen.

Publikationen mit K. Engelmann haben keinen direkten Projektbezug und sind auf die erwähnte Kooperation mit Prof. Engelmanns Augenklinik innerhalb des Klinikums Chemnitz zurückzuführen.

2.10.2 Promotionen

Im folgenden sind Promotionsarbeiten zusammengefasst bei denen es Berührungspunkte zu Projekt gab.

Abgeschlossene Dissertationen von Mitarbeitern, die mehr als 12 Monate im Projekt angestellt waren:

1. T. Kronfeld, “Aktionsfolgenbasierte Bewegungssynthese im Bereich der digitalen Fabrik,” PhD Thesis, TU Chemnitz, Chemnitz, 2019.
2. S. Kahl, “Identifying Birds by Sound: Large-scale Acoustic Event Recognition for Avian Activity Monitoring,” PhD Thesis, TU Chemnitz, Chemnitz, 2019.

Abgeschlossene Dissertationen von Mitarbeitern, die weniger als 12 Monate im Projekt angestellt waren:

3. R. Herms, “Effective Speech Features for Cognitive Load Assessment: Classification and Regression.,” PhD Thesis, TU Chemnitz, Chemnitz, 2019.

4. S. Müller, “Systematisierung und Identifizierung von Störquellen und Störerscheinungen in zeithistorischen Videodokumenten am Beispiel digitalisierter Videobestände sächsischer Lokalfernsehsender,” PhD Thesis, TU Chemnitz, Chemnitz, 2018.
5. M. Rickert, “Inhaltsbasierte Analyse und Segmentierung narrativer, audiovisueller Medien,” PhD Thesis, TU Chemnitz, Chemnitz, 2017.
6. V. Küszter, “Entwicklung eines mehrbenutzerfähigen projektionsbasierten VR-Systems und Untersuchung ausgewählter Aspekte der Nutzerinteraktion,” PhD Thesis, TU Chemnitz, Chemnitz, 2016.

Abgeschlossene Dissertationen von Doktoranden, die Berührungspunkte zum Projekt hatten, jedoch nicht darin angestellt waren:

7. B. John, “Verwendung instationärer Gasströme in der Laserfügetechnik,” PhD Thesis, TU Chemnitz, Chemnitz, 2018.
8. M. Hadzic, “Vesicle Encapsulation and Data Analysis Standardization to Characterize Large Catalytic RNAs Using Single Molecule FRET,” PhD, Universität Zürich, Zürich, 2017.
9. T. Wilhelm-Stein, “Information Retrieval in der Lehre : Unterstützung des Erwerbs von Praxiswissen zu Information Retrieval Komponenten mittels realer Experimente und Spielemechaniken,” PhD Thesis, TU Chemnitz, Chemnitz, 2016.

Laufende Dissertationen von Mitarbeitern, die mehr als 12 Monate im Projekt angestellt waren:

10. N. Englisch, “tba,” PhD Thesis, TU Chemnitz, Chemnitz, 2019.
11. T. Schlosser, “Development of a Hexagonal Deep Learning Framework,” PhD Thesis, TU Chemnitz, Chemnitz.
12. R. Manthey, “Hexagonale Bildverarbeitung,” PhD Thesis, TU Chemnitz, Chemnitz, 2019.

Nach aktuellem Stand wird Norbert Englischs Dissertation voraussichtlich in diesem Jahr eingereicht und verteidigt.

Weitere 8 Dissertationsvorhaben konnten mit den Vorarbeiten des Projekts durch Prof. Ritter an der HS Mittweida eingeworben und begonnen werden und werden Teils in Kooperation mit Prof. Eibl von der Professur Medieninformatik der TU Chemnitz durchgeführt.

2.10.3 Studentische Arbeiten

Abgeschlossene studentische Arbeiten (einschließlich Besondere Lernleistungen BELL-Arbeiten von Schülern) :

1. A. Yasin, “Dokumentation einer Funduskamera und Konzeption zur Durchführung spektralgefilterte Color-Messungen,” Pflichtpraktikum, TU Chemnitz, Chemnitz, 2019.
2. R. Wenzel, “Entwicklung und Implementierung eines robusten Archivformates mit inhaltsabhängiger, wahlfreier Fehlerrobustheit zur Langzeitarchivierung multimedialer Daten,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2019.
3. J. Tonndorf-Martini, “Implementierung eines Werkzeugs zur Analyse von AUTOSAR Projekten,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2019.
4. J. Tonndorf-Martini, “Entwicklung eines optimierten E-Learning Systems für die automotive Systemarchitektur AUTOSAR,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2019.
5. R. Tailor, “Evaluation of event-based metadata enrichment,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
6. R. Stöwesandt, “Umsetzung einer Systemkopplung zwischen der Lernplattform OPAL und einer IMS LTI Schnittstelle,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2019.
7. Y. Shaik, “Design and implementation of a web-based OCT explorer,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
8. A. Sampath Kumar, “CNN-based Audio Classification for Ambient Assisted Living and Public Transport Environments using an Extensive Combined Dataset,” Research Project, TU Chemnitz, Chemnitz, 2019.
9. M. Rößner, “Entwicklung einer Forschungsdatenbank zur Visualisierung von Patientendaten für die Unterstützung von Diagnostik und Therapie,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2019.
10. F.-T. Peters, “Extraktion und Analyse spektraler Information in der Augen-Funduskopie zur Verbesserung der Früherkennung von Augenkrankheiten oder Pathologien des Auges,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2019.
11. T.-P. Pasupuleti, “Quantitative and qualitative analysis of head poses in images,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
12. J. Ostwald, “Optimized Parameter Variation for Testing Highly Automated Driving Functions,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
13. R. Memon, “Comparison Analysis of Multiple Deep Convolutional Neural Network Models for Street-Level Object Detection,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
14. P. Kraus, “Requirement Recognition: Evaluierung von Ansätzen zur automatischen Generierung von Code aus Kundenanforderungen,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
15. M. M. Hossain, “Design and implementation of an algorithm for a collision free graphical topology definition in INVIO,” Master Thesis, TU Chemnitz, Chemnitz, 2019.

16. R. Girhotra, “Design and implementation of an Automatic Image-based Multi-Camera-Fusion System for object location in 3D environments,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
17. R. Gandhi, “Evaluation of reverse geocoding services for retrieval of location information,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
18. L. Gaitzsch, “Verwaltung und Darstellung von Testergebnissen aus AUTOSAR Projekten,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2019.
19. L. Gaitzsch, “Entwicklung eines generischen Parsers für AUTOSAR Projekte,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2019.
20. J. Dörfelt, “Intelligente Gebäudeklimatisierung auf Basis eines Sensornetzwerks und künstlicher Intelligenz,” Masterarbeit, TU Chemnitz, Chemnitz, 2019.
21. A. Bilal, “Acoustic Source Localization of Static Sources,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
22. Z. Akhtar, “Evaluation of human pose estimation algorithms from different camera perspectives,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
23. J. Varghese, “Evaluation and Design of a Low Latency Solution for Over-the-top Live Streaming,” Master Thesis, TU Chemnitz, Chemnitz, 2018.
24. J. Tonndorf-Martini, “Implementierung eines Werkzeugs zur Analyse von AUTOSAR Projekten,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2018.
25. E. Thangaraju, “Environmental Sounds Classification Using Convolutional Neural Networks,” Research Project, TU Chemnitz, Chemnitz, 2018.
26. E. Thangaraju, “Computation time evaluation of audio processing algorithms on mobile and embedded hardware,” Masterarbeit, TU Chemnitz, Chemnitz, 2018.
27. R. Siegel and R. Wenzel, “Entwicklung einer mobilen Applikation zum Messen,” Teamorientiertes Praktikum, TU Chemnitz, Chemnitz, 2018.
28. T. Schlosser, “Entwurf und Implementierung eines CHIP-basierten Systems und Demonstrators zur Aufnahme, Speicherung und Visualisierung hexagonal gerasterter Bild- und Videodaten,” Master Thesis, TU Chemnitz, Chemnitz, 2018.
29. M. Rößner and J. Lossack, “Entwicklung eines erweiterbaren Kerndatensatzes für augenmedizinische Patientendaten,” Teamorientiertes Praktikum, TU Chemnitz, Chemnitz, 2018.
30. Martin Dörfelt, “Implementierung eines Convolutional Neural Network (CNN) in Hardware am Beispiel einer Personenerkennung,” Master Thesis, TU Chemnitz, Chemnitz, 2018.
31. P. Kawczynski, “Posen- und Geometriebestimmung von Objekten im Straßenverkehr basierend auf Bildmaterial von Fischaugenkameras (Area-View) mit gegebenen, anhand künstlicher neuronaler Netze ermittelter, Region-of-Interersts,” Master Thesis, TU Chemnitz, Chemnitz, 2018.

32. A. Haddadi-Esfahani, “Real time data and video acquisition of portable ultrasound devices in ambulant point-of-care scenarios,” Master Thesis, TU Chemnitz, Chemnitz, 2018.
33. J. Götze, “Hardwarebeschleunigung von Matrixberechnungen auf Basis von GPU Verarbeitung,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2018.
34. S. Bangalore Kuppanna, “Performance Evaluation of PV Systems by modeling its energy yield using Machine learning algorithms,” Master Thesis, TU Chemnitz, Chemnitz, 2018.
35. R. Ahmed, “Design and Implementation of a web portal for handling images, metadata and analysing annotation point properties using machine learning,” Master Thesis, TU Chemnitz, Chemnitz, 2018.
36. M. Weber, “Anwendung einer automatisierten Built- und Testsystems,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2017.
37. L. Tietze, “n/a,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2017.
38. V. G. Thelkapally, “Analysis of human pose in red carpet scenarions - evaluation of stat of the art image processing methods for human pose recognition and image metadaa enrichment,” Master Thesis, TU Chemnitz, Chemnitz, 2017.
39. T. Sims, “Konzeption und Implementierung einer einfachen 3D-Engine für den Einsatz im Lehr- und Ausbildungskontext,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2017.
40. R. Siegel, “Object Localization from 3D Point Clouds,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2017.
41. T. Schlosser, “Portierung des Hexagonal Image Processing Framework CHIP zur Anwendung auf FPGA,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2017.
42. T. Schlosser, “Adaption von dcraw zur Erzeugung digitaler Bilder auf Basis des hexagonalen Rasters,” Forschungsseminar, TU Chemnitz, Chemnitz, 2017.
43. R. Mittag, “Entwicklung Statischer Analysen für AUTOSAR Steuergerätesoftware,” Masterarbeit, TU Chemnitz, Chemnitz, 2017.
44. A. Hasan, “Acoustic Source Localization,” Master Thesis, TU Chemnitz, Chemnitz, 2017.
45. E. Fabian, “Erweiterung einer SVP-Plattform um eine Akustikkomponente und deren Anwendung zur Klassikation und Lokalisierung von Geräuschquellen,” Masterarbeit, Technische Universität Chemnitz, Chemnitz, 2017.
46. H. Eskandar, “Automated Validation Of HERE Dataset Using OpenStreetMap,” Master Thesis, TU Chemnitz, Chemnitz, 2017.
47. C. Dürrling, “Entwurf und Implementierung eines Systems zur Detektion und Lokalisierung von akustischen Ereignissen im Außenbereich am Beispiel von Folgetonhörnern,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2017.

48. E. Ali Haddadi, “Performance comparison for evaluation of eyeball anterior chamber,” Hauptseminar, TU Chemnitz, Chemnitz, 2017.
49. S. Zietlow and J. Timon, “Entwicklung eines Frameworks zur Implementation von Verfahren zur Audio-Quelllokalisierung und deren Simulation,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
50. X. Xin, “n/a,” Hauptseminar, TU Chemnitz, Chemnitz, 2016.
51. F. Witscher and S. Kreutzberg, “n/a,” Teampraktikum, TU Chemnitz, Chemnitz, 2016.
52. M. Weber et al., “Umsetzung eines Werkzeuges für die Visualisierung von AUTOSAR Basissoftware,” Teampraktikum, TU Chemnitz, Chemnitz, 2016.
53. N. Vinay, “n/a,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
54. N. Vinay, “n/a,” Masterarbeit, TU Chemnitz, Chemnitz, 2016.
55. B. Trung Nguyen, R. Kumar Gade, U. Latif, P. Srivastava, and R. K. Ramakanth, “Implementation of a tool for 3D Visualization of Automotive Demonstrators,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
56. M. Seidel, “Speaker Diarization,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
57. M. Rößner, “Aufbereitung und inhaltliche Analyse klinischer Daten zum Aufbau einer ophthalmologischen Forschungsdatenbank,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2016.
58. D. Rösner, “Framework zur Verteilung, Aufzeichnung, Manipulation und Synthese von Trackingdaten für VR-Anwendungen,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
59. C. Roschke, “Entwurf und Implementierung eines webbasierten Managementsystems zur Entwicklung und Optimierung von Audio- und Videoanalysealgorithmen,” Master Thesis, Technische Universität Chemnitz, Chemnitz, 2016.
60. D.-F. Neralla, “n/a,” Masterarbeit, TU Chemnitz, Chemnitz, 2016.
61. R. Mittag, “Visualisierung von AUTOSAR Basissoftwaremodulen,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
62. M. Meinhardt Martinussen, “n/a,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
63. D. Markert, “Generische Visualisierung von Kommunikationswegen der AUTOSAR RTE,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
64. D. Markert, “Entwicklung einer generischen Testumgebung für Automotive Software Systems,” Masterarbeit, TU Chemnitz, Chemnitz, 2016.
65. B. Kindt, “n/a,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2016.
66. T. Keller, “A web-based application for data visualisation and non-linear regression analysis including error calculation for laboratory classes in natural and life sciences,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.

67. M. Heinzig, “Evaluation eines Sparsen-Hintergrundmodells im Forschungskontext der Objekterkennung,” Forschungspraktikum, Technische Universität Chemnitz, Chemnitz, 2016.
68. M. Heinzig, “Entwurf und Implementierung eines interaktiven Systems zur Evaluierung und Optimierung maschineller Lernverfahren in der virtuellen Realität,” Masterarbeit, TU Chemnitz, Chemnitz, 2016.
69. R. Greiling, “n/a,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2016.
70. A. A. G. Genaro, “n/a,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
71. M. Fiebiger, “n/a,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
72. M. Fiebiger, “n/a,” Masterarbeit, TU Chemnitz, Chemnitz, 2016.
73. T. Brösamle, “Visualisierung von AUTOSAR Applikationen auf Basis der Systembeschreibung,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2016.
74. F. Witscher, “n/a,” Hauptseminar, TU Chemnitz, Chemnitz, 2015.
75. M. Trinks, “Vergleich der Robustheit bestehender Bildqualitätsmaße im Kontext von Bilddaten,” Diplomarbeit, TU Chemnitz, Chemnitz, 2015.
76. T. Schlosser, “Entwurf und Implementierung von Optimierungs- und Funktionserweiterungen der hexagonalen Bildrasterung in der Videokompressionssoftware x264HMod,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2015.
77. D. Markert, “Anforderungsanalyse und Entwicklung einer flexiblen und modularen Benutzeroberfläche bei Sondermaschinen,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2015.
78. J. C. Lokesh Lokesh, “Design and Implementation of a Framework for Instance Search in the TRECVid Evaluation Campaign,” Masterarbeit, TU Chemnitz, Chemnitz, 2015.
79. M. Leib, “Konzeption und Implementierung eines Werkzeugs für den Test von AUTOSAR Applikationen mit Intra-ECU Kommunikation,” Masterarbeit, TU Chemnitz, Chemnitz, 2015.
80. M. Leib, “Functional description of AUTOSAR Basic Software Modules for Communication,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2015.
81. L. Krabisch, “n/a,” Hauptseminar, TU Chemnitz, Chemnitz, 2015.
82. B. Kindt, R. Greiling, and S. Conrad, “Hexagonale Bildtransformation,” Teampraktikum, TU Chemnitz, Chemnitz, 2015.
83. T. Keller, “Webbasierte Curve Fitting Anwendungen,” Hauptseminar, TU Chemnitz, Chemnitz, 2015.
84. F. Hänchen, “Tracing von Signalen in AUTOSAR Systems,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2015.
85. F. Hänchen, “Prototypische Entwicklung einer generischen Health-Monitoring-Architektur für AUTOSAR-Systeme,” Masterarbeit, TU Chemnitz, Chemnitz, 2015.

86. J. Gehre, “Entwicklung eines 3D-Scanners aus Raspberry Pi und Micro-Projektor,” BELL, TU Chemnitz, Chemnitz, 2015.
87. M. Fiebinger, “n/a,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2015.
88. S. Conrad, “Erstellung und Evaluation von Testdatensätzen zur Leistungsdatenermittlung multimedialer Datenverarbeitungssysteme,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2015.
89. J. Albert, “n/a,” BELL, TU Chemnitz, Chemnitz, 2015.
90. M. Heinzig, “Entwurf und Implementierung eines erweiterbaren Systems zur Detektion von Objekten in großen Beständen audiovisueller Medien,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2014.

In der Projektlaufzeit begonnene, laufende studentische Arbeiten:

1. M. Teich, “Visualisierung, Geo- und Ereignisdaten,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2019.
2. T. Schröder, “Hexagonal Image Processing - Structure Image Processing and Point-Spread-Function simulations/Entwurf und Implementierung von Textursegmentierungs- und klassifizierungsverfahren für hexagonale Bildrasterung,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
3. M.-U. Sabir, “People counting based on head detection in crowded scenes (with surveillance environment),” Master Thesis, TU Chemnitz, Chemnitz, 2019.
4. M. Friedrich, “Binäres Template Matching auf Gabor Bildern zur Fehleranalyse in der Halbleiterindustrie,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2019.
5. W. Farooq, “Edge detection and Contrast Enhancement of Wafers using Collinearity,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
6. S. D. Ramkumar, “Evaluation of multiple bicycle detection algorithms or implementation of a multiple bicylce detection algorithm,” Research Project, TU Chemnitz, Chemnitz, 2019.
7. T. Poornachandra, “Face Recognition for Chemnitz University of Technology relevant persons,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
8. S. Mejri, “Entwicklung und Evaluation automatisierter Text-Strukturierung von augenmedizinischen Textdaten Fundus,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
9. S. Mejri, “Evaluation automatisierter Text-Aufbereitungsmethoden von augenmedizinischen Patientendaten,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
10. T. Keller, “Evaluation und Weiterentwicklung einer Webanwendung zur Regressionsanalyse auf Basis von Nutzerstudien,” Master Thesis, TU Chemnitz, Chemnitz, 2019.
11. S. Conrad, “Evaluation of synthesis of hexagonal images - Povray vs. alternative approaches,” Master Thesis, TU Chemnitz, Chemnitz, 2019.

12. N. Udas, “Concept Detection in Medical Images Using Xception Models,” Master Thesis, TU Chemnitz, Chemnitz, 2018.
13. R. Stöwesandt, “Umsetzung einer Systemkopplung zwischen der Lernplattform OPAL und einer IMS LTI Schnittstelle,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2018.
14. S. Shakil, “OCT and/or Fundus Image Classification using CNNs,” Research Project, TU Chemnitz, Chemnitz, 2018.
15. Y. Obaid, “Evaluation of head pose algorithms,” Master Thesis, TU Chemnitz, Chemnitz, 2018.
16. H. Munir, “Bicycle Detection/Counting,” Master Thesis, TU Chemnitz, Chemnitz, 2018.
17. D. Madhiyazhagan, “Automatic classification of fundus imaging and Diabetic Retinopathy specific pathologies,” Pflichtpraktikum, TU Chemnitz, Chemnitz, 2018.
18. S. Conrad, “Szenenrenderung mit POVRay unter Verwendung hexagonaler Pixel,” Forschungspraktikum, 2018.
19. J. Tonndorf-Martini, “Entwicklung eines optimierten E-Learning Systems für die automotive Systemarchitektur AUTOSAR,” Bachelorarbeit, TU Chemnitz, Chemnitz, 2017.
20. M. Seidel, “Akustische Ereignisdetektion, TrecVid,” Master Thesis, TU Chemnitz, Chemnitz, 2017.
21. Kindt, Benjamin, “Blender animiertes Rendern zur Ground Truth Generierung,” Forschungspraktikum, TU Chemnitz, Chemnitz, 2017.

Literatur

[1] Focusrite OctoPre MKII. Available on `https://us.focusrite.com/mic-pres/octopre-mkii` (2016), last accessed at 27. October 2016.

[2] Focusrite OctoPre MKII Dynamic. Available on `http://us.focusrite.com/mic-pres/octopre-mkii-dynamic` (2016), last accessed at 27. October 2016.

[3] Genelec Loudspeaker. Available on `https://www.genelec.com/` (2016), last accessed at 1st. November 2016.

[4] HDSPe MADICard PCIe. Available on `https://www.rme-audio.de/en/products/hdsp_madi.php` (2016), last accessed at 1st. November 2016.

[5] Microphone Audio Technica AT4040. Available on `http://www.audio-technica.com/cms/wired_mics/9b6aac05c5aca887/` (2016), last accessed at 28. October 2016.

[6] Microphone JustIn JM-714 Clipmic. Available on http://www.justmusic.de/de-de/pa-licht/mikrofone/instrumentenmikrofone/150811/jm-714-clipmic.html (2016), last accessed at 28. October 2016.

[7] Microphone MXL 840. Available on http://www.mxlmics.com/microphones/800-series/840-pair/ (2016), last accessed at 27. October 2016.

[8] Microphone NowSonic Calibration. Available on http://nowsonic.com/en/products/studio/calibration/ (2016), last accessed at 27. October 2016.

[9] Microphone Rode NT5-MP. Available on http://www.rode.com/microphones/nt5 (2016), last accessed at 28. October 2016.

[10] Presonus HP4 - Kopfhörerverstärker. Available on https://www.presonus.com/products/hp4 (2016), last accessed at 10. April 2017.

[11] RME ADI-648. Available on https://www.rme-audio.de/en/products/adi_648.php (2016), last accessed at 28. October 2016.

[12] Roland - A-88 | MIDI Keyboard Controller. Available on https://www.roland.com/de/products/a-88/ (2016), last accessed at 10. April 2017.

[13] Rosendahl Nano Clock. Available on http://www.rosendahl-studiotechnik.de/nanoclocks.html (2016), last accessed at 1st. November 2016.

[14] Yamaha DGX-660 WH Keyboard. Available on https://de.yamaha.com/de/products/musical_instruments/pianos/portable_grand/dgx-660/?mode=model (2016), last accessed at 10. April 2017.

[15] BLX188/CVL Dual Channel Lavalier Wireless System. Available on https://www.shure.com/americas/products/wireless-systems/blx-wireless-systems/blx188-cvl (2017), last accessed at 10. April 2017.

[16] BLX288/PG58 Dual Channel Handheld Wireless System. Available on https://www.shure.com/americas/products/wireless-systems/blx-wireless-systems/blx288-pg58 (2017), last accessed at 10. April 2017.

[17] TANNOY Loudspeaker. Available on http://www.tannoy.com/creation/ (2017), last accessed at 10. April 2017.

[18] CNN based head pose estimator (Jan 2020), https://github.com/ZhuangleiScut

[19] Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3908–3916. IEEE (jun 2015), http://ieeexplore.ieee.org/document/7299016/

[20] Ahmed, R.: Design and Implementation of a web portal for handling images, metadata and analysing annotation point properties using machine learning. Master's thesis, TU Chemnitz, Chemnitz (Nov 2018)

[21] Akhtar, Z.: Evaluation of human pose estimation algorithms from different camera perspectives. Master's thesis, TU Chemnitz, Chemnitz (Sep 2019)

[22] Antonelli, M., Gibaldi, A., Beuth, F., Duran, A.J., Canessa, A., Chessa, M., Hamker, F.H., Chinellato, E., Sabatini, S.P.: A hierarchical system for a distributed representation of the peripersonal space of a humanoid robot. IEEE Transactions on Autonomous Mental Development 6(4), 259–273 (2014), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6844843

[23] Berger, A., Eibl, M., Heinrich, S., Herms, R., Kahl, S., Knauf, R., Kurze, A., Rickert, M., Ritter, M.: ValidAX - Validierung der Frameworks AMOPA und XTRIEVAL. Schlussbericht CSR-15-01, Chemnitz (Feb 2015), https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa-158977

[24] Berger, A., Eibl, M., Heinrich, S., Knauf, R., Kürsten, J., Kurze, A., Rickert, M., Ritter, M.: sachsMedia - Cooperative Producing, Storage, Retrieval and Distribution of Audiovisual Media (FKZ: 03IP608). Schlussbericht CSR-12-04, Chemnitz (Sep 2012), https://www.bibliothek.tu-chemnitz.de/uni_biblio/frontdoor.php?source_opus=12988

[25] Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., Pinto, H.P.d.O., Raiman, J., Salimans, T., Schlatter, J., Schneider, J., Sidor, S., Sutskever, I., Tang, J., Wolski, F., Zhang, S.: Dota 2 with Large Scale Deep Reinforcement Learning. arXiv preprint arXiv:1912.06680. (2019), http://arxiv.org/abs/1912.06680

[26] Beuth, F.: Visual attention in primates and for machines - neuronal mechanisms. Ph.D. thesis, Technische Universität Chemnitz (In Press)

[27] Beuth, F., Hamker, F.H.: A mechanistic cortical microcircuit of attention for amplification, normalization and suppression. Vision Research 116(Part B), 241–257 (2015), http://www.ncbi.nlm.nih.gov/pubmed/25883048

[28] Bilal, A.: Acoustic Source Localization of Static Sources. Master's thesis, TU Chemnitz, Chemnitz (Feb 2019)

[29] Bonastre, J.F., Wils, F., Meignier, S.: ALIZE, a Free Toolkit for Speaker Recognition. pp. 737–740. http://dx.doi.org/10.1109/icassp.2005.1415219

[30] Börner, R., Kowerko, D., Hadzic, M.C.A.S., König, S.L.B., Ritter, M., Sigel, R.K.O.: Simulations of camera-based single-molecule fluorescence experiments. PLOS ONE 13(4), 1–23 (Apr 2018), https://doi.org/10.1371/journal.pone.0195277

[31] Börner, R., Kowerko, D., Miserachs, H.G., Schaffer, M.F., Sigel, R.K.: Metal ion induced heterogeneity in RNA folding studied by smFRET. Coordination Chemistry Reviews 327–328, 123–142 (Nov 2016), http://linkinghub.elsevier.com/retrieve/pii/S0010854516300601

[32] Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1812.08008 [cs] (May 2019), http://arxiv.org/abs/1812.08008, arXiv: 1812.08008

[33] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime Multi-Person 2d Pose Estimation using Part Affinity Fields. CoRR abs/1611.08050 (2016), http://arxiv.org/abs/1611.08050

[34] Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., Rosen, D.B., et al.: Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Transactions on neural networks 3(5), 698–713 (1992)

[35] Carpenter, G.a., Martens, S., Ogas, O.J.: Self-organizing information fusion and hierarchical knowledge discovery: A new framework using ARTMAP neural networks. Neural Netw 18, 287–295 (2005)

[36] Cheon, S., Lee, H., Kim, C.O., Lee, S.H.: Convolutional Neural Network for Wafer Surface Defect Classification and the Detection of Unknown Defect Class. IEEE Transactions on Semiconductor Manufacturing 32(2), 163–170 (may 2019), https://ieeexplore.ieee.org/document/8657760/

[37] De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C.O., Raine, R., Hughes, J., Sim, D.A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P.T., Suleyman, M., Cornebise, J., Keane, P.A., Ronneberger, O.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine 24(9), 1342–1350 (Sep 2018), http://www.nature.com/articles/s41591-018-0107-6

[38] Dörfelt, J.: Intelligente Gebäudeklimatisierung auf Basis eines Sensornetzwerks und künstlicher Intelligenz. Masterarbeit, TU Chemnitz, Chemnitz (2019)

[39] Dürrling, C.: Entwurf und Implementierung eines Systems zur Detektion und Lokalisierung von akustischen Ereignissen im Außenbereich am Beispiel von Folgetonhörnern. Bachelorarbeit, TU Chemnitz, Chemnitz (Nov 2017)

[40] Erler, R., Manthey, R., Hussein, H., Siegel, R., Kowerko, D.: Realisation of an Audio & Video Laboratory for Precise Object Localisation and Tracking. In: Elektronische Sprachsignalverarbeitung 2018 (ESSV 2018) (Mar 2018), http://essv2018.de/wp-content/uploads/2018/03/36_Manthey_ESSV2018.pdf

[41] Fiscus, J., Joy, D., Michel, M., Awad, G., Smeaton, A.F., Jones, G.J.F., Kraaij, W., Quenot, G., Ritter, M., Eskevich, M., Ordelman, R., Huet, B., Larson, M.: TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking (2016)

[42] Gandhi, R.: Evaluation of reverse geocoding services for retrieval of location information. Master's thesis, TU Chemnitz, Chemnitz (Feb 2019)

[43] Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017. pp. 776–780 (2017), https://doi.org/10.1109/ICASSP.2017.7952261

[44] Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron. https://github.com/facebookresearch/detectron (2018)

[45] Goeau, H., Kahl, S., Glotin, H., Planque, R., Joly, A.: Overview of BirdCLEF 2018: monospecies vs. soundscape bird identification. In: CLEF 2018 Working Notes. p. 12 (2018), ceur-ws.org/Vol-2125/invited_paper_9.pdf

[46] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. In: 3rd International Conference on Learning Representations (ICLR 2015). pp. 1–11 (dec 2014), http://arxiv.org/abs/1412.6572

[47] Goëau, H., Glotin, H., Planqu\'e, R., Vellinga, W.P., Joly, A.: LifeCLEF Bird Identification Task 2017. In: CLEF working notes 2017 (2017), http://www.imageclef.org/lifeclef/2017/bird

[48] Grossberg, S.: Adaptive Resonance Theory: how a brain learns to consciously attend, learn, and recognize a changing world. Neural Netw 37, 1–47 (jan 2013), http://www.ncbi.nlm.nih.gov/pubmed/23149242

[49] Group, V.Q.E., others: Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II (FR_tv2). ftp://ftp. its. bldrdoc.

gov/dist/ituvidq/Boulder_VQEG_jan_04/VQEG_PhaseII_FRTV_Final_Report_SG9060E.doc, 2003 (2003), http://ci.nii.ac.jp/naid/10015195696/

[50] Gu, S., Holly, E., Lillicrap, T., Levine, S.: Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In: Proceedings - IEEE International Conference on Robotics and Automation (ICRA). pp. 3389–3396. IEEE (2017)

[51] Guobing, Y.: Head pose estimation (Jan 2020), https://github.com/yinguobing/head-pose-estimation

[52] Hadzic, M., Börner, R., König, S.L.B., Kowerko, D., Sigel, R.K.O.: Reliable State Identification and State Transition Detection of Fluorescence Intensity-Based smFRET Data. Journal of Physical Chemistry B p. in revision (2018)

[53] Hamker, F.H.: The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. Cerebral Cortex 15(4), 431–447 (2004), http://www.ncbi.nlm.nih.gov/pubmed/15749987

[54] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (2016), https://www.cv-foundation.org/openaccess/content{_}cvpr{_}2016/html/He{_}Deep{_}Residual{_}Learning{_}CVPR{_}2016{_}paper.html

[55] Herms, R.: Effective Speech Features for Cognitive Load Assessment: Classification and Regression. PhD Thesis, TU Chemnitz, Chemnitz (2019), urn:nbn:de:bsz:ch1-qucosa2-333464

[56] Hussein, H., Ritter, M., Manthey, R., Schloßhauer, J., Fabian, E., Heinzig, M.: Acoustic Event Classification for Ambient Assisted Living and Healthcare Environments. In: Proc. of 27 Konferenz Elektronische Sprachsignalverarbeitung (ESSV). Leipzig, Germany (März 2016)

[57] Jamalian, A., Beuth, F., Hamker, F.H.: The performance of a biologically plausible model of visual attention to localize objects in a virtual reality. In: Proceedings of the International Conference on Artificial Neural Networks. pp. 447–454 (2016)

[58] John, B.: Verwendung instationärer Gasströme in der Laserfügetechnik. PhD Thesis, TU Chemnitz, Chemnitz (Jul 2018), https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa2-312405

[59] John, B., Markert, D., Englisch, N., Grimm, M., Ritter, M., Hardt, W., Kowerko, D.: Quantification of geometric properties of the melting zone in laser-assisted wel-

ding. In: Proceedings of Lasers in Manufacturing 2017. pp. 1–9. München (Jun 2017), http://www.wlt.de/lim/

[60] Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Lombardo, J.C., Planqué, R., Palazzo, S., Müller, H.: LifeCLEF 2017 Lab Overview: Multimedia Species Identification Challenges. In: Jones, G.J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction, vol. 10456, pp. 255–274. Springer International Publishing, Cham (2017), http://link.springer.com/10.1007/978-3-319-65813-1_24

[61] Jones, J.P., Palmer, L.A.: An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58(6), 1233–58 (Dec 1987), http://www.ncbi.nlm.nih.gov/pubmed/3437332

[62] Kahl, S.: The 2018 lifeclef bird identification task baseline system, https://github.com/kahst/BirdCLEF-Baseline, [Online: zuletzt zugegriffen 25.01.2020].

[63] Kahl, S.: Source code of the tucmi submission to birdclef2017, https://github.com/kahst/BirdCLEF2017, [Online: zuletzt zugegriffen 25.01.2020].

[64] Kahl, S.: Identifying Birds by Sound: Large-scale Acoustic Event Recognition for Avian Activity Monitoring. PhD Thesis, TU Chemnitz, Chemnitz (2019)

[65] Kahl, S., Hussein, H., Fabian, E., Schloßhauer, J., Thangaraju, E., Kowerko, D., Eibl, M.: Acoustic Event Classification Using Convolutional Neural Networks. In: INFORMATIK 2017. pp. 2177–2188. Gesellschaft für Informatik, Bonn, Chemnitz (2017), https://dl.gi.de/handle/20.500.12116/3989

[66] Kahl, S., Richter, D., Roschke, C., Rickert, M., Heinzig, M., Kowerko, D., Eibl, M., Ritter, M.: Technische Universität Chemnitz and Hochschule Mittweida at TRECVID Instance Search 2017. TRECVid Workshop Proceedings 2017, 1–7 (Sep 2017)

[67] Kahl, S., Roschke, C., Rickert, M., Richter, D., Zywitz, A., Hussein, H., Manthey, R., Heinzig, M., Kowerko, D., Eibl, M., Ritter, M.: Technische Universitat Chemnitz at TRECVID Instance Search 2016 (Sep 2016), https://www.researchgate.net/publication/312211838_Technische_Universitat_Chemnitz_at_TRECVID_Instance_Search_2016

[68] Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M., Eibl, M.: Large-Scale Bird Sound Classification using Convolutional Neural Networks. In: CEUR Workshop Proceedings (Working Notes of CLEF 2017 - Conference and Labs of the Evaluation). vol. 1866 (Sep 2017), ceur-ws.org/Vol-1866/paper_143.pdf

[69] Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., Eibl, M.: A Baseline for Large-Scale Bird Species Identification in Field Recordings. In: Working notes of CLEF (2018), http://ceur-ws.org/Vol-2125/paper_85.pdf

[70] Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., Eibl, M.: Recognizing Birds from Sound - The 2018 BirdCLEF Baseline System. CoRR abs/1804.07177 (2018), http://arxiv.org/abs/1804.07177

[71] Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874 (June 2014)

[72] Kowerko, D.: 3D Indoor Audio Localization of Moving Objects (Jul 2019)

[73] Kowerko, D., Manthey, R., Heinz, M., Kronfeld, T., Brunnett, G.: Fast and accurate creation of annotated head pose image test beds as prerequisite for training neural networks. In: INFORMATIK 2017. pp. 2221–2229. Gesellschaft für Informatik, Bonn, Chemnitz (2017), https://dl.gi.de/handle/20.500.12116/3994

[74] Kowerko, D., Richter, D., Heinzig, M., Kahl, S., Helmert, S., Brunnett, G.: Evaluation of CNN-based algorithms for human pose analysis of persons in red carpet scenarios. In: INFORMATIK 2017. pp. 2201–2209. Gesellschaft für Informatik, Bonn, Chemnitz (2017), https://dl.gi.de/handle/20.500.12116/3991

[75] Kretzschmar, T., Kowerko, D.: Image related metadata generation, storage and retrieval for big datasets. In: Chemnitzer Informatik Berichte 2020. vol. CSR-20-01, pp. 9–28. Universitätsbibliothek/Universitätsverlag, Chemnitz (Jan 2020)

[76] Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Adv Neural Inf Process Syst 25 - NIPS 2012, pp. 1097–1105 (2013)

[77] Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. In: 5th International Conference on Learning Representations (ICLR 2017). pp. 1–14 (jul 2016), http://arxiv.org/abs/1607.02533

[78] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015), http://www.nature.com/doifinder/10.1038/nature14539

[79] Lee, K.B., Cheon, S., Kim, C.O.: A Convolutional Neural Network for Fault Classification and Diagnosis in Semiconductor Manufacturing Processes. IEEE Transactions on Semiconductor Manufacturing 30(2), 135–142 (2017), http://ieeexplore.ieee.org/document/7867863/

[80] Lee, K.B., Kim, C.O.: Recurrent feature-incorporated convolutional neural network for virtual metrology of the chemical mechanical planarization process. Journal of Intelligent Manufacturing pp. 1–14 (2018), http://link.springer.com/10.1007/s10845-018-1437-4

[81] Lewke, D., Dohnke, K.O., Zühlke, H.U., Cerezuela Barreto, M., Schellenberger, M., Bauer, A., Ryssel, H.: Thermal laser separation – a novel dicing technology fulfilling the demands of volume manufacturing of 4H-SiC devices. In: Materials Science Forum. vol. 821, pp. 528–532 (2015)

[82] Lin, T.Y., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. CoRR abs/1405.0312 (2014), http://arxiv.org/abs/1405.0312

[83] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single Shot MultiBox Detector. arXiv:1512.02325 [cs] 9905, 21–37 (2016), http://arxiv.org/abs/1512.02325, arXiv: 1512.02325

[84] Manthey, R.: Algorithm development and evaluation with virtual environments. TRECVID Workshop Proceedings (2017)

[85] Manthey, R., Kowerko, D.: Hexagonal image generation by virtual multi-grid-camera. In: Karwowski, W., Ahram, T. (eds.) Intelligent Human Systems Integration 2019. AISC, vol. 903, pp. 17–22. Springer International Publishing, San Diego, CA, USA (01 2019)

[86] Manthey, R., Schmidsberger, F., Thomanek, R., Roschke, C., Rolletschke, T., Platte, B., Ritter, M., Kowerko, D.: An Exploratory Inspection of the Detection Quality of Pose and Object Detection Systems by Synthetic Data. In: Stephanidis, C. (ed.) HCI International 2019 - Posters. pp. 287–294. Springer International Publishing, Cham (2019)

[87] Manthey, R., Thomanek, R., Roschke, C., Ritter, M., Kowerko, D.: Synthetic ground truth generation for testing, technology evaluation and verification (synttev). In: Proceedings of the 32nd International BCS Human Computer Interaction Conference (HCI 2018) (07 2018)

[88] Manthey, R., Thomanek, R., Roschke, C., Rolletschke, T., Platte, B., Ritter, M., Kowerko, D.: Visual system examination using synthetic scenarios. In: Karwowski, W., Ahram, T. (eds.) International Conference on Intelligent Human Systems Integration. AISC, vol. 903, pp. 418–422. Springer International Publishing, San Diego, CA, USA (01 2019)

[89] Meignier, S., Merlin, T.: LIUM SpkDiarization: an Open Source Toolkit for Diarization. In: CMU SPUD Workshop. Dallas (Texas, USA) (2010)

[90] Meng, W., Xiao, W.: Energy-Based Acoustic Source Localization Methods: A Survey. Sensors 17(2) (2017), http://www.mdpi.com/1424-8220/17/2/376

[91] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature 518(7540), 529–533 (2015), http://dx.doi.org/10.1038/nature14236

[92] Nakazawa, T., Kulkarni, D.V.: Wafer map defect pattern classification and image retrieval using convolutional neural network. IEEE Transactions on Semiconductor Manufacturing 31(2), 309–314 (2018), https://ieeexplore.ieee.org/document/8263132

[93] Pasupuleti, T.P.: Quantitative and qualitative analysis of head poses in images. Master's thesis, TU Chemnitz, Chemnitz (Jul 2019)

[94] Piczak, K.J.: Esc-50: Dataset for environmental sound classification, https://github.com/karoldvl/ESC-50, [Online: zuletzt zugegriffen 18.04.2018].

[95] Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: 25th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2015, Boston, MA, USA, September 17-20, 2015. pp. 1–6 (2015), https://doi.org/10.1109/MLSP.2015.7324337

[96] Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. CoRR abs/1612.08242 (2016), http://arxiv.org/abs/1612.08242

[97] Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement. arXiv preprint arXiv:1804.02767 (apr 2018), http://arxiv.org/abs/1804.02767

[98] Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. CoRR abs/1506.01497 (2015), http://arxiv.org/abs/1506.01497

[99] Rickert, M.: Inhaltsbasierte Analyse und Segmentierung narrativer, audiovisueller Medien. PhD Thesis, TU Chemnitz, Chemnitz (2017)

[100] Ritter, M.: Optimierung von Algorithmen zur Videoanalyse: ein Analyseframework für die Anforderungen lokaler Fernsehsender. No. Bd. 3 in Wissenschaftliche Schriftenreihe Dissertationen der Medieninformatik, Univ.-Verl, Chemnitz (2014)

[101] Ritter, M., Kowerko, D., Hussein, H., Schlosser, T., Heinzig, M., Manthey, R., Bahr, G.S.: Simplifying Accessibility Without Data Loss: An Exploratory Study on Object Preserving Keyframe Culling (Feb 2016)

[102] Ritter, M., Kowerko, D., Manthey, R., John, B., Grimm, M.: Quantifizierung der geometrischen Eigenschaften von Schmelzzonen bei Laserschweißprozessen. In: Forum Bildverarbeitung. Karlsruhe, Germany (2016)

[103] Roschke, C.: Entwurf und Implementierung eines webbasierten Managementsystems zur Entwicklung und Optimierung von Audio- und Videoanalysealgorithmen. Master's thesis, Technische Universität Chemnitz, Chemnitz, Germany (2016)

[104] Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., Meignier, S.: An Open-Source State-of-the-Art Toolbox for Broadcast News Diarization. In: INTERSPEECH (August 2013)

[105] Russell, S.J., Norvig, P.: Künstliche Intelligenz: Ein moderner Ansatz. Pearson, Higher Education, München, 3., aktualisierte auflage edn. (2012)

[106] Salamon, J., Jacoby, C., Bello, J.: A dataset and taxonomy for urban sound research. In: Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014 (11 2014)

[107] Sampath Kumar, A.: CNN-based Audio Classification for Ambient Assisted Living and Public Transport Environments using an Extensive Combined Dataset. Research Project, TU Chemnitz, Chemnitz (May 2019)

[108] Sampath Kumar, A., Erler, R., Kowerko, D.: A Real-Time Demo for Acoustic Event Classification in Ambient Assisted Living Contexts. In: Proceedings of the 27th ACM International Conference on Multimedia - MM '19. pp. 2205–2207. ACM Press, Nice, France (2019), http://dl.acm.org/citation.cfm?doid=3343031.3350600

[109] Sampath-Kumar, A., Erler, R., Kowerko, D.: CNN-based Audio Classification for Environmental Sounds, Ambient Assisted Living and Public Transport Environments using an Extensive Combined Datas. In: Chemnitzer Informatik Berichte 2020. vol. CSR-20-01, pp. 29–66. Universitätsbibliothek/Universitätsverlag, Chemnitz (Jan 2020)

[110] Schlosser, T., Beuth, F., Friedrich, M., Kowerko, D.: A Novel Visual Fault Detection and Classification System for Semiconductor Manufacturing Using Stacked Hybrid Convolutional Neural Networks. In: 2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). pp. 1511–1514. IEEE, Zaragoza, Spain (Sep 2019), https://ieeexplore.ieee.org/document/8869311/

[111] Schlosser, T., Beuth, F., Friedrich, M., Kowerko, D.: Fehlerdetektion und -klassifikation bei Laserschneidprozessen mittels Deep Neural Networks. In: Chemnitzer Informatik Berichte 2020. vol. CSR-20-01, pp. 67–81. Universitätsbibliothek/Universitätsverlag, Chemnitz (Jan 2020)

[112] Scholkopf, Bernhard Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge, MA, USA (2001)

[113] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. Nature 529(7587), 484–489 (2016), http://dx.doi.org/10.1038/nature16961

[114] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., Hassabis, D.: Mastering the game of Go without human knowledge. Nature 550(7676), 354–359 (2017), http://dx.doi.org/10.1038/nature24270

[115] Sobral, A.: BGSLibrary: An opencv c++ background subtraction library. In: IX Workshop de Visao Computacional (WVC'2013). vol. 7 (2013), http://www.academia.edu/download/32708022/andrews_WVC2013.pdf

[116] Steffen, F.D., Khier, M., Kowerko, D., Cunha, R.A., Börner, R., Sigel, R.K.O.: Metal ions and sugar puckering balance single-molecule kinetic heterogeneity in RNA and DNA tertiary contacts. Nature Communications 11(1), 104 (Dec 2020), http://www.nature.com/articles/s41467-019-13683-4

[117] Steinberg Cubase: Steinberg Cubase Pro 8.5 ED, https://www.steinberg.net/de/home.html, last accessed at 18. March 2017

[118] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: International conference on artificial neural networks (ICANN 2018). pp. 270–279 (2018)

[119] Taubert, S., Mauermann, M., Kahl, S., Kowerko, D., Eibl, M.: Species Prediction based on Environmental Variables using Machine Learning Techniques. In: Working notes of CLEF (2018), http://ceur-ws.org/Vol-2125/paper_93.pdf

[120] Thangaraju, E.: Computation time evaluation of audio processing algorithms on mobile and embedded hardware. Masterarbeit, TU Chemnitz, Chemnitz (Sep 2018)

[121] Thomanek, R., Roschke, C., Manthey, R., Platte, B., Rolletschke, T., Heinzig, M., Vodel, M., Kowerko, D., Kahl, S., Zimmer, F., Eibl, M., Ritter, M.: University of applied sciences mittweida and chemnitz university of technology at trecvid 2018. TRECVID Workshop Proceedings (2018)

[122] Thomanek, R., Roschke, C., Platte, B., Manthey, R., Rolletschke, T., Heinzig, M., Vodel, M., Zimmer, F., Eibl, M., Ritter, M.: A scalable system architecture for activity detection with simple heuristics. In: Winter Applications of Computer Vision Workshops (WACVW). pp. 27–34. Winter Applications of Computer Vision, IEEE, Waikoloa Village, Hawaii, USA (01 2019)

[123] Thomanek, R., Roschke, C., Platte, B., Rolletschke, T., Schlosser, T., Heinzig, M., Kowerko, D., Vodel, M., Zimmer, F., Eibl, M., Ritter, M.: University of Applied Sciences Mittweida and Chemnitz University of Technology at TRECVID Instance Search 2019. In: TRECvid Workshop Proceedings 2019. p. 9. Gaithersburg, UNITED STATES (Nov 2019), https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/HSMW_TUC_ins.pdf

[124] Thomanek, R., Roschke, C., Platte, B., Rolletschke, T., Schlosser, T., Heinzig, M., Vodel, M., Kowerko, D., Zimmer, F., Eibl, M., Ritter, M.: University of Applied Sciences Mittweida and Chemnitz University of Technology at TRECVID ActEv 2019. In: TRECvid Workshop Proceedings 2019. p. 7. Gaithersburg, UNITED STATES (Nov 2019), https://www-nlpir.nist.gov/projects/tvpubs/tv19.papers/HSMW_TUC_actev.pdf

[125] Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional Pose Machines. CoRR abs/1602.00134 (2016), http://arxiv.org/abs/1602.00134

[126] Wilton, P.: MADI (Multichannel Audio Digital Interface). In: Audio Engineering Society Conference: UK 3rd Conference: AES/EBU Interface. pp. EBU–14 (September 1989), http://www.aes.org/e-lib/browse.cfm?elib=5340

[127] Xu, R., Wunsch, D.: Clustering. John Wiley & Sons (2008)

[128] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Proc European Conf Computer Vision 2014 - ECCV 2014. pp. 1–11 (2014), http://link.springer.com/chapter/10.1007/978-3-319-10590-1{_}53

[129] Zietlow, T., Hussein, H., Kowerko, D.: Acoustic Source Localization in Home Environments - The Effect of Microphone Array Geometry. In: Elektronische Sprachsignalverarbeitung 2017 (ESSV 2017), pp. 219–226. No. 86 in Studientexte zur Sprachkommunikation, TUDpress, Saarbrücken (2017), http://essv2017.coli.uni-saarland.de/index.html

[130] Zühlke, H.U., Eberhardt, G., Ullmann, R.: TLS-Dicing – an innovative alternative to known technologies. In: ASCM 2009. pp. 28–32 (2009)

Chemnitzer Informatik-Berichte

In der Reihe der Chemnitzer Informatik-Berichte sind folgende Berichte erschienen:

CSR-13-01 Navchaa Tserendorj, Uranchimeg Tudevdagva, Ariane Heller, Grenzgänger - Integration of Learning Management System into University-level Teaching and Learning, Januar 2013, Chemnitz

CSR-13-02 Thomas Reichel, Gudula Rünger, Multi-Criteria Decision Support for Manufacturing Process Chains, März 2013, Chemnitz

CSR-13-03 Haibin Xu,Thomas Reichel, Gudula Rünger, Michael Schwind, Softwaretechnische Verknüpfung der interaktiven Softwareplattform Energy Navigator und der Virtual Reality Control Platform, Juli 2013, Chemnitz

CSR-13-04 International Summerworkshop Computer Science 2013, Proceedings of International Summerworkshop 17.7. - 19.7.2013, Juli 2013, Chemnitz

CSR-13-05 Jens Lang, Gudula Rünger, Paul Stöcker, Dynamische Simulationskopplung von Simulink-Modellen durch einen Functional-Mock-up-Interface- Exportfilter, August 2013, Chemnitz

CSR-14-01 International Summerschool Computer Science 2014, Proceedings of Summerschool 7.7.-13.7.2014, Juni 2014, Chemnitz

CSR-15-01 Arne Berger, Maximilian Eibl, Stephan Heinich, Robert Herms, Stefan Kahl, Jens Kürsten, Albrecht Kurze, Robert Manthey, Markus Rickert, Marc Ritter, ValidAX - Validierung der Frameworks AMOPA und XTRIEVAL, Januar 2015, Chemnitz

CSR-15-02 Maximilian Speicher, What is Usability? A Characterization based on ISO 9241-11 and ISO/IEC 25010, Januar 2015, Chemnitz

CSR-16-01 Maxim Bakaev, Martin Gaedke, Sebastian Heil, Kansei Engineering Experimental Research with University Websites, April 2016, Chemnitz

Chemnitzer Informatik-Berichte

CSR-18-01 Jan-Philipp Heinrich, Carsten Neise, Andreas Müller, Ähnlichkeitsmessung von ausgewählten Datentypen in Datenbanksystemen zur Berechnung des Grades der Anonymisierung, Februar 2018, Chemnitz

CSR-18-02 Liang Zhang, Guido Brunnett, Efficient Dynamic Alignment of Motions, Februar 2018, Chemnitz

CSR-18-03 Guido Brunnett, Maximilian Eibl, Fred Hamker, Peter Ohler, Peter Protzel, StayCentered - Methodenbasis eines Assistenzsystems für Centerlotsen (MACeLot) Schlussbericht, November 2018, Chemnitz

CSR-19-01 Johannes Dörfelt, Wolfram Hardt, Christian Rosjat, Intelligente Gebäudeklimatisierung auf Basis eines Sensornetzwerks und künstlicher Intelligenz, Februar 2019, Chemnitz

CSR-19-02 Martin Springwald, Wolfram Hardt, Entwicklung einer RAD-Plattform im Kontext verteilter Systeme, März 2019, Chemnitz

CSR-19-03 André Böhle, René Schmidt, Wolfram Hardt, Evaluation von Signaleigenschaften zur Lokalisierung von Einschlägen mit Piezokeramischen Sensoren, März 2019, Chemnitz

CSR-19-04 Johannes Götze, René Schmidt, Wolfram Hardt, Hardwarebeschleunigung von Matrixberechnungen auf Basis von GPU Verarbeitung, März 2019, Chemnitz

CSR-19-05 Vincent Kühn, Reda Harradi, Wolfram Hardt, Expert System for Adaptive Flight Missions, Juni 2019, Chemnitz

CSR-19-06 Samer Salamah, Guido Brunnett, Christian Mitschke, Tobias Heß, Synthesszing gait motions from spline-based progression functions of controlled shape, Juni 2019, Chemnitz

CSR-19-07 Martin Eisoldt, Carsten Neise, Andreas Müller, Analyse verschiedener Distanzmetriken zur Messung des Anonymisierungsgrades θ, Juni 2019, Chemnitz

CSR-19-08 André Langer, Valentin Siegert, Martin Gaedke, Informationsverwertung basierend auf qualitätsoptimierten semistrukturierten Datenbeständen im Wachstumskern “LEDS”, Juli 2019, Chemnitz

CSR-20-01 Danny Kowerko, Chemnitzer Linux-Tage 2019 - LocalizeIT Workshop, Januar 2020, Chemnitz

CSR-20-02 Robert Manthey, Tom Kretzschmar, Falk Schmidsberger, Hussein Hussein, René Erler, Tobias Schlosser, Frederik Beuth, Marcel Heinz, Thomas Kronfeld, Maximilian Eibl, Marc Ritter, Danny Kowerko, Schlussbericht zum InnoProfile-Transfer Begleitprojekt localizeIT, Januar 2020, Chemnitz

Chemnitzer Informatik-Berichte

ISSN 0947-5125

Herausgeber: Fakultät für Informatik, TU Chemnitz
Straße der Nationen 62, D-09111 Chemnitz