Digitale Menschliche Zwillinge (DiMenZ)

Gereon KORTENBRUCK, Lukas JAKUBCZYK, Dennis ROLOFF

PROLAB Produkt+Produktion,
Technische Hochschule Georg Agricola
Herner Straße 45, D-44787 Bochum

Kurzfassung: In dieser Studie wird die Einsatzeignung von State-of-the-Art Deep Learning-Applikationen zur Erkennung menschlicher Bewegungen sowie zu deren Spiegelung in den digitalen Raum vorgestellt. Die Grundlage hierfür bildet die Taxonomie der Bewegungen nach MTM. Die Restriktion ist, dass ausschließlich markerlose Verfahren zum Einsatz kommen.

Schlüsselwörter: Deep Learning, Human Pose Estimation, MTM, Ergonomiebewertung, digitaler menschlicher Zwilling

1. Einleitung

Die Erkennung von menschlichen Körperhaltungen während der Verrichtung einer Tätigkeit ist im Zuge ergonomischer Betrachtung sehr wichtig und wird derzeitig entweder über Motion-Capture-Systeme oder deterministische digitale Menschmodelle durchgeführt. Die Verwendung von neuronalen Netzwerken für diese Aufgabe findet nicht statt, obwohl die Erkennung von Menschen und deren Bewegungen durch ebensolche Netzwerke in den vergangenen zwei Jahren große Fortschritte gemacht hat. So ist die Einschätzung der Lage markanter Punkte des menschlichen Körpers und deren anatomisch richtige Verbindung(en) untereinander mit Hilfe CN- Netzwerken (convolutional neural networks) im 2D aber auch 3D-Bereich und in Echtzeit möglich. Neben der Verwendung dieser Technologien zur Schaffung von digitalen menschlichen Zwillingen, um diese mit Echtzeitdaten der Produktions-umgebung zu koppeln und so eine Fabrikation im Sinne des Industrie 4.0- Gedankens zu etablieren, sind mannigfaltige Einsatzmöglichkeiten denkbar.

2. Die Grundmethoden der neuronalen Netzwerke

Künstliche neuronale Netze entstammen aus den Neurowissenschaften als mathematisches Modell für die Dynamik von Nervenzellen im Gehirn. Sie bestehen aus Knoten die durch Kanten miteinander verbunden sind. Die Knoten sind in Schichten angelegt und im einfachsten Fall so untereinander verknüpft, dass jeder Knoten (Neuron) mit allen Knoten der benachbarten Schichten verbunden ist. Die Verbindungen (Kanten) sind gewichtet. Die Berechnung erfolgt schichtweise von der Eingangsschicht bis zur Ausgangsschicht. Der Wert eines jeden Eingangsknotens, beispielsweise der Helligkeitswert eines Bild-Pixels, wird mit dem Kantengewicht multipliziert und dient zusammen mit den anderen Produkten (restliche Knoten der Eingangsschicht, die zu diesem Knoten führen und deren Gewichte) als Eingang. Dort werden die Produkte aufsummiert. Eine Aktivierungsfunktion wird auf diese Summe angewendet und das Ergebnis ist der Wert dieses Knotens. Diese Berechnung setzt sich so lange fort, bis die Ausgangsschicht erreicht ist. Das Ergebnis ist ein Vorher-

sagewert für einen bekannten Wert. Es wird die Differenz gemessen, anhand welcher die Gewichte aller Kanten neuinitialisiert werden. Das Verfahren nennt sich Backpropagation und besteht im Wesentlichen auf der Anwendung der Kettenregeln auf mehrdimensionale partielle Ableitungen. Die Ergebnisse der Vorhersage werden solange mit bekannten Daten verglichen, bis die Differenz (Verlustfunktion) minimal ist. Das Netz hat dann das den Eingangsdaten zugrundeliegende Muster gelernt resp. die Vorhersagefunktion approximiert und sollte für unbekannte Daten eine gute Vorhersage treffen. Die Güte dieser Vorhersage hängt von der Güte und Größe des Datensatzes, der für die Lernaufgabe verwendet wurde, ab. Zu erwähnen ist, dass die Eingabeschicht perfekt an die zu verarbeitenden Daten angepasst ist; sollen durch die Vorhersagefunktion etwa Objekte in Schwarz-Weiß-Bildern klassifiziert werden, so wird die Bildgröße in Pixeln vorgegeben und die Eingangsschicht hat dann entsprechend die Bildpixelbreite multipliziert mit der Bildpixelhöhe als Anzahl der Neuronen. Dies zeigt, dass bei visuellen Daten die Anzahl der Neuronen sehr schnell sehr groß werden kann. Daher wurde für diese Anwendung die Netzwerkarchitektur Convolutional Neural Network (CNN); zu Deutsch Faltungsnetzwerke, entwickelt. Ein CNN extrahiert automatisch visuelle oder, im Allgemeinen, räumliche Merkmale einer Eingabe, so dass es in der Lage ist, eine optimale Vorhersage basierend auf diesen Merkmalen zu tätigen. Es enthält folgende Bausteine: Eingabeschicht (Matrix aus Bildhöhe [px] multipliziert mit der Bildbreite [px] für jeweils einen der drei Farbkanäle), Faltungs- und Bündelungsschicht (wird nachfolgend darauf eingegangen) sowie der vollständig verbundenen Schicht, die bereits erklärt wurde. Der Hauptbaustein von CNNs ist die Faltungsschicht (Convolutional Layer). Die Faltung ist ein mathematisches Verfahren um zwei Informationssätze zusammenzuführen. In diesem Fall wird die Faltung mit einem Convolution Filter angewendet, um eine so genannte Feature Map zu erzeugen.

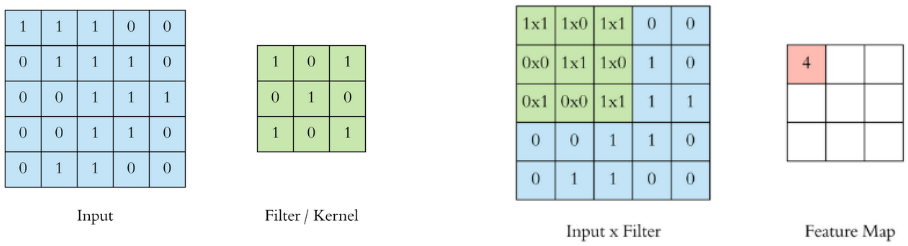

Abbildung 1: *Beispiel für ein Eingangsbild und einen Scanvorgang mit Darstellung des ersten Filtervorgangs*

In der Abbildung 1 ist auf der linken Seite ein Input, in diesem Fall ein Eingangsbild, eines Convolutional Layer zu sehen. Auf der rechten Seite ist ein Convolution Filter (Kernel) abgebildet. Dieses Kernel wird genutzt um den Input zu scannen. An jedem Punkt wird eine Matrixmultiplikation durchgeführt und die entstandenen Ergebnisse werden summiert. Diese Summe wird in der Feature Map eingetragen. Es werden mehrere Faltungsvorgänge, mit jeweils unterschiedlichen Filtern an einem Eingang durchgeführt, um unterschiedliche Merkmale eines Bildes zu extrahieren. So etwa der Sobel-Operator, der in einem Bild die Bereiche findet, in denen eine starke Änderung der Intensität oder der Farbe vorliegt. Diese Bereiche stellen die Kanten eines abgebildeten Objektes dar. (vgl. Fagna, 2018) Nachdem der komplette Convolution Vorgang abgeschlossen ist, wird zur Reduzierung der Dimensionalität eine Bündelung, das sogenannte Pooling durchgeführt. Pooling-Layer verkleinern die Höhe und Breite der einzelnen Feature Maps unabhängig voneinander. Die Tiefe bleibt jedoch bestehen. (vgl. Dertat, 2017) Im Gegensatz zum Convolution Vorgang besitzt das

Pooling keine Parameter und wählt im Falle des Max Poolings immer den größten Wert der Eingabe im Pooling-Fenster aus. In Abbildung 2 ist das Fenster 2x2 groß und es wird in Zweierschritten gescannt. Jede Farbe zeigt ein Pooling-Fenster. Da die Fenstergröße 2 ist und dieses Fenster immer 2 Schritte „springt“, gibt es keine Überlappungen der einzelnen Fenster. Des Weiteren zeigt die Abbildung, dass die Schritt- und Fensterkonfiguration unter Beibehaltung wichtiger Informationen die Feature Map halbiert.

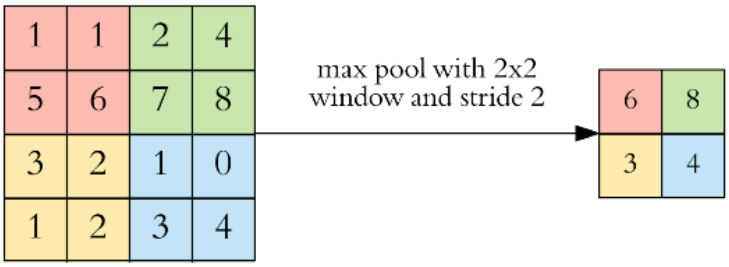

Abbildung 2: *Beispiel eines Max-Poolings*

Durch die Halbierung der Höhe und Breite, wurde die Anzahl der Gewichte auf ¼ des Inputs reduziert. Nach den Convolution- und Pooling-Vorgängen folgen die vollständig verbundenen Schichten, um die Architektur des Convolutional Neural Network abzuschließen. Die Ausgabe der vollständig verbundenen Schichten ist das Ergebnis des Netzes. (vgl. Dertat, 2017) Bei Bilddateien ist es die Einteilung in eine bestimmte Klasse. Liegen etwa zwei Klassen vor, z.B. Hund und Katze, so kann das Netz bei einem neuen Bild erkennen, ob es sich bei dem Objekt um eine Katze oder einen Hund handelt. Eine weitere sehr wichtige Architektur sind die Rekurrenten Neuronalen Netze (RNN). Im Gegensatz zu den beiden bereits vorgestellten Architekturen, wo die Berechnung von der Eingabe- zur Ausgabeschicht schrittweise erfolgt, sind bei einem RNN Verbindungen zwischen Neuronen der aktuellen mit Neuronen der gleichen Schicht, wie aber auch mit Neuronen vergangener Schichten etabliert, wodurch das Netz eine Art Gedächtnis erhält. Dadurch kann es gut Muster in Daten erkennen, die einer zeitlichen Veränderung unterliegen. Zum Beispiel menschliche Bewegungen.

3. Die Grundmethoden der Erkennung menschlicher Körperhaltungen (Human Pose Estimation)

Ziel von Human Pose Estimation Programmen ist es, auf Basis von Bildmaterial die Lage markanter Körperpunkte (Keypoints) und deren anatomisch richtige Verbindung untereinander abzuschätzen. Das geschieht über Faltungsnetzwerke, Rekurrente Netzwerke oder Netzwerke, die eine Mischung der beiden zuvor beschriebenen Netzwerkarchitekturen darstellen. Aktuelle Verfahren lassen sich auf zwei Grundmethoden, die Top-Down und Bottom-Up Methode, zurückführen. Bei der Top-Down Methode wird im ersten Schritt um jede Person im Bild ein Rechteck (Bounding Box) gezeichnet. Die Problematik der Mehrpersonenerkennung wird somit auf die Erkennung einer einzelnen Person innerhalb jeder ermittelten Bounding Box heruntergebrochen (vgl. Cheng et al., 2019). Anwendungen nach der Top-Down Methode sind in der Regel wenig anfällig gegenüber Skalierungsunterschieden verschiedener Perso-

nen, jedoch aufgrund der separaten Haltungserkennung jeder einzelnen Person sehr rechenaufwendig (vgl. Cheng et al., 2019). Ein weiterer Nachteil dieser Methode ist die Problematik der Mehrfachdetektion, dem mehrmaligen erkennen der gleichen Person (Vgl. Fang et al., 2018). Nach dem Prinzip der Bottom-Up Methode werden zunächst auf Basis sogenannter Heatmaps alle Keypoints innerhalb eines Bildes, ohne Zuordnung zu einer bestimmten Person, ermittelt. Anschließend werden die ermittelten Keypoints einzelnen Personen zugeordnet und anatomisch korrekt verknüpft (vgl. Cheng et al., 2019). Somit sind Anwendungen nach der Bottom-Up Methode schneller als solche nach der Top-Down Methode, jedoch verlieren diese mit abnehmender Bildauflösung stark an Genauigkeit (vgl. Cheng et al., 2019).

4. Anwendung zur Klassifizierung menschlicher Bewegungen nach der MTM-Taxonomie

Basierend auf den Ergebnissen der Programme, arbeitet das PROLAB Produkt+ Produktion an einer Auswertungsmethode, die die Ausgabe von Deep Learning-Applikationen in MTM-Bewegungen umwandelt. Für die Anwendbarkeit dieser Methode ist die Ausgabe der x- und y-Koordinaten eines jeden Keypoints unabdingbar. Die Bewegung eines Probanden wird Bild für Bild untersucht. Für jedes Einzelbild wird nebst einem digitalen Skelett, welches auf den Körper der Person gemappt wird, eine Datei mit den besagten Koordinaten erzeugt. Mit Hilfe dieser Dateien kann für jeden Keypoint die Positionsdifferenz von Bild zu Bild ermittelt und für weitere Auswertungen genutzt werden. Hierzu muss der Inhalt der Dateien in Blöcke unterteilt werden. Jeder Block stellt einen Keypoint auf dem Körper dar. Die erste Zahl des Blocks zeigt die x-Koordinate, die zweite die y-Koordinate und die dritte Zahl den Einschätzungswert, dass sich der Keypoint auf der genauen Körperstelle befindet. Für die Differenzbildung der so vorbereiteten Koordinaten wird eine dafür angepasste Excel Mappe verwendet.

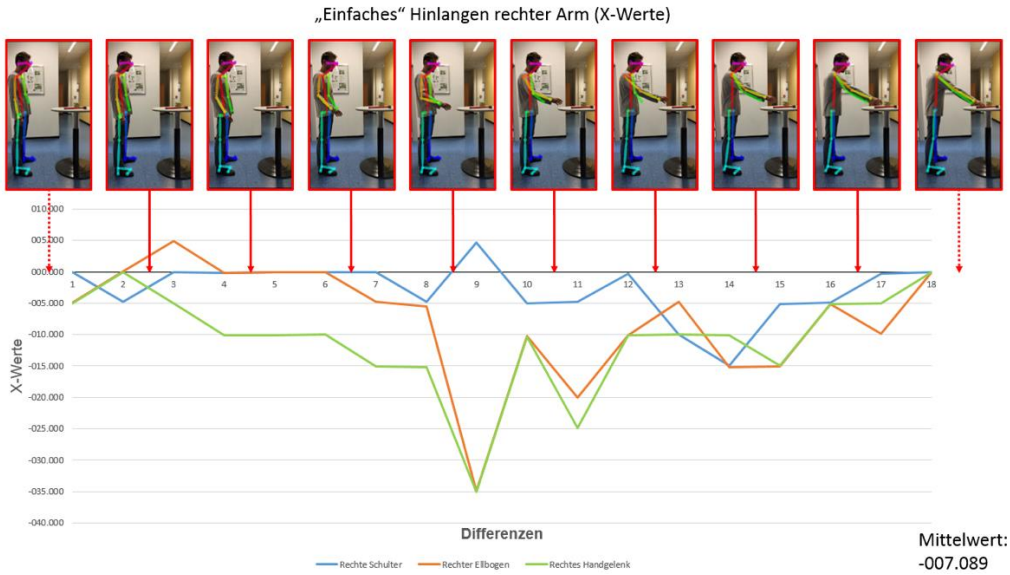

Abbildung 3: *Übersicht über die Verteilung der Keypoints über die gesamte "Hinlang"-Bewegung*

Bei der Betrachtung der einzelnen Werte einer „Hinlang“-Bewegung wird aufgrund der großen Differenzen deutlich, dass die Keypoints der Arme und der Hände relevant für die Auswertung sind. Ein Beispiel für das Ergebnis der Differenzbildung einer

„Hinlang“-Bewegung wird in Abbildung 3 dargestellt. In diesem Fall hat ein Proband eine „Hinlang“-Bewegung ausgeführt. Der Proband hat an einer bestimmten Stelle ein Minimum. Infolge dessen ist der Mittelwert aller „Hinlang“-Bewegungen negativ, sodass prognostiziert werden kann, dass jede „Hinlang“-Bewegung anhand des negativen Mittelwerts erkannt werden kann. Momentaner Stand der Forschung am PROLAB Produkt+Produktion ist, dass für die an MTM-2 angelehnten Bewegungen, die Anzahl der Aufnahmen pro Person und Betrachtungsposition stark erhöht (n>100) wird, um eine statistische valide Aussage zu haben, welche Keypoints mit welchen Mittelwerten eine bestimmte Bewegung repräsentieren. Zudem wird ein eigner Datensatz zur Validierung unserer Methode im Sinne des CRISP-Standards (Cross Industry Standard Process for Data Mining) entwickelt. Dieser Datensatz trägt den Namen „Bewegungen des täglichen Lebens“ und zeigt verschieden Personen in unterschiedlichen Umgebungen bei der Verrichtung alltäglicher Aufgaben. Für die beschriebene Aufgabe kamen folgende Deep-Learning-Programme zum Einsatz: AlphaPose, OpenPose, Monocular 3D Human Pose Estimator, PifPaf, Pose2Seg und Stacked Hourglass Networks for Human Pose Estimation. Diese Programme unterscheiden sich stark voneinander. Die durchgeführte Nutzwertanalyse, die neben dem Hauptkriterium, der Ausgabe der Koordinaten der Schlüsselpunkte, Faktoren wie Lizenz (in einem Fall 100.000 $ p.a. für kommerzielle Zwecke), Hardwareanforderungen (in einem Fall 2 High-End-Grafikkarten) oder Installierbarkeit (keine der Anwendung unter 2 Stunden für geschultes Personal) betrachtete, grenzte das Feld ein. Die drei, in den egeren Kreis gezogenen Systeme sind mit einem Ausgabebild in der Abbildung 4 aufgeführt.

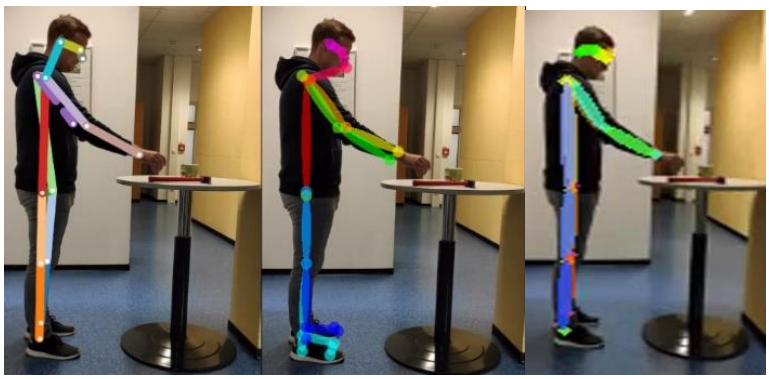

Abbildung 4: *Simulation einer Extremsituation mit verdecktem linken Arm mit [links] PifPaf, [mitte] OpenPose und [rechts] AlphaPose*

5. Anwendung zur Spiegelung menschlicher Bewegungen in den digitalen Raum

Die Ausgabe der Lage markanter Körperpunkte als Höhen- und Breitenkoordinaten eines vom System aus initialisierten Koordinatensystems kann theoretisch dazu benutzt werden, den Menschen in eine 3D-Umgebung zu spiegeln. Die Idee dahinter ist, dass diese sogenannten digitalen menschlichen Zwillinge mit den ebenfalls in dieser Umgebung befindlichen Zwillingen einer Produktionsumgebung (Maschinen, Werkzeuge, Einzelteile, Halbfabrikate, Produkte) interagieren. Die Kopplung von menschlichen und maschinellen Daten kann einen transparenten und nachverfolgbaren Fabrikationsprozess im Sinne des Industrie 4.0-Gedankens darstellen. Ein solcher Prozess kann mit Algorhitmen des maschinellen Lernens optimiert werden. Ein

gut vorstellbares Szenario wird nachfolgend beschrieben. Während einer wertschöpfenden Tätigkeit wird der Mensch im digitalen Raum gespiegelt. Dort wird er durch ein computererzeugtes Konstrukt, dem so genannten Agenten beobachtet. Dieser verfügt über die gleiche digitale Anatomie und versucht mittels des Nachahmungslernens die Bewegungen des Menschen zu imitieren. Einmal angelernt, kann dieser Agent vor eine ähnliche Aufgabe gestellt werden. Um diese zu lösen wird dieser Agent mehrfach geklont. Mit der Methode von Versuch-und-Irrtum und der aus der Verhaltensbiologie bekannten Konditionierung durch „Belohnung“ und „Bestrafung“ können die Klone eine Lösung für die neue Aufgabe finden. Diese Vorgehensweise wird „bestärkendes Lernen“ genannt und die Lösung der Aufgabe heißt Policy. Eine Policy gibt an, welche Aktion bei einer beliebigen Observation aus der Lernumgebung ausgeführt werden soll, um die Belohnung zu maximieren. Die Belohnung kann auch negativ sein. So könnten ergonomisch ungünstige Bewegungen, die zur Lösung der Geamtaufgabe führen, „bestraft“ werden und ergonomisch günstige „belohnt“ werden. Das Ergebnis ist eine computergenerierte Bewegungsanalyse für eine neue Produktvariante. Die Vision ist, dass bereits während der Konstruktion eines neuen Produktes die dazugehörigen manuellen Arbeitsschritte mitsamt einer ergonomischen Bewertung bereitgestellt werden. Im PROLAB Produkt+Produktion haben die ersten Arbeiten an der Erzeugung der Klone für sehr einfache Tätigkeiten begonnen. Die Ergebnisse sind vielversprechend, zeigen aber auch gleichzeitig Forschungsbedarf auf.

6. Ausblick

Die Erkennung menschlicher Bewegungen nach einer an die MTM-2 Systematik angelehnten Taxonomie funktioniert und wird an einem neuen und diversen Datensatz validiert. Die anschließende Integration in ein Ergonomie-Bewertungssystem lässt sich verhältnismäßig leicht realisieren. Die Erzeugung digitaler Zwillinge bedarf weiterer Forschung bezüglich des Hauptmodells zur Erkennung der menschlichen Bewegungen im 3 D-Bereich ohne spezielle Kameras und im Bereich der Fusionierung mehrerer Modelle (Körperhaltung, Objekterkennung, menschliche Sprache). Das PROLAB Produkt+Produktion arbeitet im Rahmen von Forschungsarbeiten daran, ein „offenes“ System für die Echtzeiterfassung valider Daten für die Zeitermittlung und Ergonomiebewertung zu entwickeln. Neben dem dargestellten Schwerpunkt ergeben sich aus dieser Arbeit heraus weitere Anwendungsfälle, die über das produzierende Umfeld hinausgehen.

7. Literaturverzeichnis

Cheng B, Xiao B, Wang J, Shi H, Huang T, Zhang L (2019) Bottom-up Higher-Resolution Networks for Multi-Person Pose Estimation. Abgerufen am 17.11.2019 von https://paperswithcode.com/paper/bottom-up-higher-resolution-netorks-for

Dertat, Arden (2017): Applied Deep Learning - Part 4: Convolutional Neural Networks. Abgerufen am 09.12.2019 von https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Fagna, Ashish (2018): Understanding Edge Detection (Sobel Operator). Abgerufen am 09.12.2019 von https://medium.com/datadriveninvestor/understanding-edge-detection-sobel-operator-2aada303b900

Fang H, Xie S, Tai Y, Lu C (2018) RMPE: Regional Multi-Person Pose Estimation. Abgerufen am 05.12.2019 von http://openaccess.thecvf.com/content_iccv_2017/html/Fang_RMPE_Regional_Multi-Person_ICCV_2017_paper.html

Gesellschaft für
Arbeitswissenschaft e.V.

Digitale Arbeit, digitaler Wandel, digitaler Mensch?

66. Kongress der
Gesellschaft für Arbeitswissenschaft

TU Berlin
Fachgebiet Mensch-Maschine-Systeme

HU Berlin
Professur Ingenieurpsychologie

16. – 18. März 2020, Berlin

GfA-Press

Bericht zum 66. Arbeitswissenschaftlichen Kongress vom 16. – 18. März 2020

TU Berlin, Fachgebiet Mensch-Maschine-Systeme
HU Berlin, Professur Ingenieurpsychologie

Herausgegeben von der Gesellschaft für Arbeitswissenschaft e.V.
Dortmund: GfA-Press, 2020
ISBN 978-3-936804-27-0

NE: Gesellschaft für Arbeitswissenschaft: Jahresdokumentation

Als Manuskript zusammengestellt. Diese Jahresdokumentation ist nur in der Geschäftsstelle erhältlich.

Schriftleitung: Matthias Jäger
im Auftrag der Gesellschaft für Arbeitswissenschaft e.V.

Screen design und Umsetzung
© 2020 fröse multimedia, Frank Fröse
office@internetkundenservice.de · www.internetkundenservice.de