Computer-Vision-basierte Aktivitätserkennung bei Schweinen

Lukas Hesse [1], Maik Fruhner [2], Heiko Tapken [3] und Henning Müller[4]

Abstract: Die Sicherstellung des Tierwohls ist einer der Kernaspekte in der modernen Nutztierhaltung. Da sich durch den steigenden Bedarf an Lebensmitteln und dem steigenden Kostendruck immer mehr Landwirte dazu gezwungen sehen, immer größere Tierzahlen zu halten, fällt es vor dem Hintergrund des Fachkräftemangels schwieriger, diesem Aspekt nachzukommen. Aus diesem Grund müssen Technologien zur Unterstützung von Landwirten entwickelt werden, welche datenbezogene hochwertige Entscheidungshilfen geben können. Einen solchen Ansatz erarbeitet das Team des Forschungsprojektes SmartTail, bei dem unter anderem eine Computer-Vision-basierte Aktivitätserkennung erarbeitet wird. Durch die nicht-invasive und kostengünstige Hardware können so potenziell flächendeckend Systeme zur Unterstützung der Landwirte implementiert werden. Innerhalb dieser Arbeit wird sich mit der videobasierten Aktivitätserkennung bei Schweinen beschäftigt. Besonders betrachtet wird dabei das Problem des Schwanzbeißens. Dieses ist in der Schweinehaltung bekannt, aber aufgrund der multifaktoriellen Ursachen existiert bisher weder ein System zur Vorhersage noch zum Erkennen solcher Attacken. Aus diesem Grund werden innerhalb dieser Arbeit mehrere state-of-the-art Modelle zur bildbasierten Aktivitätserkennung betrachtet und miteinander verglichen, um so ein effektives System zur Aktivitätserkennung bei Schweinen zu entwickeln.

Keywords: Computer Vision, Aktivitätserkennung, Künstliche Intelligenz, Mastschweine

1 Einleitung

Ein in der Schweinemast immer wieder auftretendes Problem, zu dessen eindeutigen Ursachen nur wenige Informationen vorliegen, ist das Schwanzbeißen. Dies ist eine Verhaltensstörung bei Schweinen, bei denen sich die Tiere gegenseitig die Ringelschwänze verletzen. Dies kann zu entzündeten Wunden und einem stagnierenden Wachstum führen. Zudem breitet sich dieses Verhalten nach einmaligem Auftreten schnell auf weitere Tiere der Gruppe aus, weshalb eine frühzeitige Erkennung oder sogar eine Vorwarnung von größter Bedeutsamkeit wäre.

[1] HS Osnabrück, Fakultät IuI, Albrechtstr. 30, 49076 Osnabrück, lukas.hesse@hs-osnabrueck.de, https://orcid.org/0000-0001-5247-7538

[2] HS Osnabrück, Fakultät IuI, Albrechtstr. 30, 49076 Osnabrück, m.fruhner@hs-osnabrueck.de, https://orcid.org/0000-0002-9094-6996

[3] HS Osnabrück, Fakultät IuI, Albrechtstr. 30, 49076 Osnabrück, h.tapken@hs-osnabrueck.de, https://orcid.org/0000-0002-0685-5072

[4] Hof Fleming, Ehrener Kirchweg 6, 49624 Löningen, henning.mueller@hof-fleming.de

Aktuell wird eine solche Attacke erst während der regelmäßigen Bonitur erkannt, indem ein lädierter Schwanz festgestellt wird. Zu diesem Zeitpunkt ist allerdings nur noch das Opfer der Attacke identifizierbar und der Aggressor kann nicht von der Gruppe getrennt werden. Durch die multifaktoriellen Ursachen solcher Beißattacken existiert bis heute kein Vorhersagesystem und die Informationen zum eigentlichen Beißvorgang sind minimal.

Da keine Methode zur sicheren Eindämmung des Schwanzbeißens besteht, wird in der konventionellen Tierhaltung oft auf das Kupieren zurückgegriffen. Hierbei werden den Ferkeln die Ringelschwänze abgeschnitten, bevor andere Artgenossen diese abbeißen können. Dieses Vorgehen wird allerdings scharf von Tierschützern kritisiert und der „Aktionsplan Kupierverzicht“ [AkKu18] plant eine schrittweise Eindämmung dieser Praktik.

Ausgehend von diesem Problem wurde die von einem Landwirt geleitete Projektgruppe SmartTail gegründet, um ein technisches System zur automatischen Erkennung von Schwanzbeißen zu entwickeln. Durch ein solches System soll zum einen dem Landwirt das Auftreten von Schwanzbeißen in Echtzeit mitgeteilt werden, zum anderen soll es dabei helfen, die existierende Datenlücke für den Vorgang des Schwanzbeißens zu schließen. Der im Projekt gewählte Lösungsansatz beruht auf einer videobasierten Aktivitätserkennung, bei der durch eine dauerhafte Überwachung sowohl Aggressor und Opfer als auch das Auftreten von Schwanzbeißen erkannt werden sollen.

Um einen weiteren Schritt zur Ursachenforschung des Schwanzbeißens zu erreichen, werden verschiedene Deep-Learning-Architekturen erarbeitet, um eine generelle Aktivitätserkennung bei Schweinen zu erreichen. Hierfür wurde in einem ersten Schritt ein umfangreicher Datensatz geschaffen, bevor verschiedene state-of-the-art Architekturen auf unterschiedlich detaillierten Aktivitätskatalogen trainiert wurden.

2 Stand der Forschung

Videobasierte Ansätze zur Aktivitätserkennung von Schweinen sind ein aktuelles Thema in der Forschung. Eine Vielzahl von vergleichbaren Arbeiten nutzen Deckenkameras, um eine möglichst gute Übersicht über einzelne Buchten zu erhalten [Na19]. Zudem kann ein Fokus auf Computer-Vision-basierte Ansätze festgestellt werden, was mit der generellen Transition im Feld der Aktivitätserkennung einherzugehen scheint. Bisherige Arbeiten zu diesem Thema lassen sich meist an zwei Faktoren unterscheiden. Der Erste bezieht sich auf die erkannten Aktivitäten. Da es bisher keine allgemeingültige Auflistung von Schweineaktivitäten gibt, werden auch bei Arbeiten mit einem ähnlichen Erkennungsziel verschiedene Aktivitätskataloge gewählt. Der zweite Faktor verweist auf die genutzten Techniken. Obwohl eine allgemeine Verschiebung hin zu Computer Vision mit neuronalen Netzen basierten Ansätzen erkennbar ist, gibt es keinen Konsens zur optimalen Architektur.

So untersuchen beispielsweise Zheng et al. [Zh18] die Körperhaltung von Schweinen mithilfe von Tiefenkamerabildern und einem Faster R-CNN Modell. Analog zu diesem

Ansatz untersuchen Nasirahmadi et al. [Na19] die Performance verschiedener Objektdetektoren zur Erkennung von drei verschiedenen Körperhaltungen. In einer Arbeit von Alameer et al. [AKB20] wurde sowohl mit YOLO- als auch mit Faster R-CNN-Detektoren gearbeitet, um komplexere Aktivitäten zu erkennen.

In verschiedenen Arbeiten von Yang et al. [Ya18] [Ya20] wurden ähnlich komplexe Aktivitäten untersucht. Hierbei wurde sich jedoch nicht auf einfache Objektdetektoren verlassen. Stattdessen wurde eine einfache Two-Stream-Architektur genutzt, bei der ein separater Input zur Analyse des optischen Flusses angewandt wird. Auf diese Weise werden in zwei Streams jeweils die zeitlichen und örtlichen Merkmale verarbeitet.

Andere Arbeiten nutzen bereits modernere Architekturen, welche die direkte Verarbeitung von Videos erlauben. So nutzen Zhang et al. [Hu20] ein I3D-Modell zur Erkennung von fünf verschiedenen Aktivitäten, während auf dem gleichen Datensatz eine SlowFast-Architektur von Li et al. [Zh20] trainiert wurde. In einer Arbeit von Chen et al. [Ch20] wurde eine abgewandelte Two-Stream Architektur mit einem CNN- und einem RNN-Pfad zur binären Erkennung von aggressivem Verhalten genutzt.

Zum speziellen Ziel der spatio-temporalen Erkennung (ST-Erkennung) liegen bisher wenige Vergleichsarbeiten vor. In einem Versuch von Liu et al. [Li20] wird ein Netzwerk zur Erkennung von Schwanzbeißen trainiert. Dieses ermöglicht eine binäre Klassifikation, wobei jeweils Paare von Schweinen betrachtet werden.

Generell ist in den hier beachteten Arbeiten ein Trend hin zu komplexeren End-to-End Modellen erkennbar, wie beispielsweise dem I3D [CaZi17] oder dem Slowfast-Netzwerk [Fe19]. Allerdings liegen speziell für die ST-Erkennung kaum Arbeiten vor und für eine solche Aktivitätserkennung in Kombination mit Multi-Object Erkennung sind keine Arbeiten bekannt. Bei der Verfolgung des Trends ist ein solcher Ansatz jedoch der logische nächste Schritt, welcher bessere Ergebnisse verspricht.

3 Versuchsaufbau

Der in dieser Arbeit genutzte Datensatz beruht auf dauerhaften Videoaufnahmen einzelner Buchten, welche innerhalb des Projektes SmartTail aufgenommen wurden. Hierzu wurden in verschiedenen Ställen insgesamt fünf Buchten mit IP-Kameras überwacht. Diese Kameras verfügen sowohl über einen Farb- als auch einen aktiven Infrarotmodus und sind mit einem orthogonalen Blickwinkel unter der Stalldecke angebracht, sodass durch eine Kamera eine gesamte Bucht abgedeckt wird. Die für dieses Paper genutzten Kameras besitzen eine Auflösung von 3840 * 2160 Pixeln und nehmen mit einer Framerate von 30 FPS auf, wobei eine leichte Fish-Eye Verzerrung existiert.

Durch diesen Aufbau liegen projektintern inzwischen mehrere komplette Mastzyklen vor, was einem rohen Datenmaterial von mehr als 15 TB entspricht. Während eines Zyklus kann die Anzahl der Schweine pro Bucht variieren, wobei jedoch bei einem Großteil der

Aufnahmen 13 Schweine pro Bucht zu sehen sind. Dies entspricht der maximalen Auslastung einer Bucht ihrer Größe unter Berücksichtigung der rechtlichen Vorgaben.

4 Methodik

In diesem Kapitel werden sowohl der für dieses Paper gewählte Ansatz zur Datenvorverarbeitung als auch die genutzten künstlichen neuronalen Netzwerke, welche zum Vergleich herangezogen wurden, vorgestellt. Hierbei wird auf die Besonderheiten dieser Netzstrukturen hingewiesen und ihre Vor- und Nachteile werden aufgezeigt.

4.1 Datensatz

Bei einem Vergleich verwandter Arbeiten zur automatisierten Erkennung von Schweineaktivitäten fällt auf, dass bei diesen die genutzten Ethogramme zur Aktivitätsdefinition oft bewusst klein gewählt sind, um für den jeweiligen Use Case die besten Ergebnisse zu erhalten. Bei dem hier angegangenen Problem geht es jedoch um eine generelle Erkennung von Aktivitäten, weshalb die verwandten Arbeiten nicht die benötigte genaue Unterteilung von Aktivitäten mitbringen.

Aus diesem Grund wurde auf ein Ethogramm aus der lange etablierten Schweineforschung zurückgegriffen. In der Arbeit von Zonderland et al. [Zo11] geht es im Besonderen um die ausgeführten Aktivitäten von Schweinen vor einer Schwanzbeißattacke. Innerhalb dieser Arbeit wurden 33 verschiedene Aktivitäten definiert, welche ein Schwein zu jeder Zeit ausführen kann.

Ein weiterer Vorteil dieses Ethogramms ist die bereits existierende Unterteilung dieser feingranularen Aktivitäten in allgemeinere Oberkategorien. So existieren beispielsweise fünf übergreifende Klassen: „Posture“, „Performed Behavioural States“, „Received Behavioural States“, „Performed Behavioural Events“ und „Received Behavioural Events“.

Diese umfassen die Körperhaltung eines Tieres, welche zu jedem Zeitpunkt zugewiesen werden kann. Zusätzlich werden hier längere ausgeführte oder aufgezwungene Verhaltenszustände festgelegt. Ein Tier befindet sich immer in mindestens einem ausführenden Zustand und kann zu jedem Zeitpunkt gleichzeitig auch Teil einer aufgezwungenen Aktivität sein. Hierzu zählen beispielsweise das Untersuchen von Spielzeug in der Bucht, zu dem es keine aufgezwungene Aktivität gibt, als auch das Manipulieren eines anderen Schweins, zu dem eine zugehörige aufgezwungene Aktion existiert.

Zusätzlich zu den länger andauernden Aktivitätsklassen existieren die ausgeführten und aufgezwungenen Events. Diese sind im Gegensatz zu Zuständen lediglich kurzweilig und müssen nicht zu jeder Zeit von einem Tier ausgeführt werden. In diese Kategorie fällt beispielsweise das Schwanzbeißen und das Gebissenwerden. Zu diesen fünf

Oberkategorien existieren teilweise weitere Unterkategorien, bevor die einzelnen elementaren Aktivitäten definiert werden. Diese hierarchische Unterteilung ist bei den hier angestellten Untersuchungen besonders hilfreich, da zu Beginn nicht bekannt ist, wie genau ein automatisiertes System die einzelnen von Schweinen ausgeführten Aktivitäten unterscheiden kann. Auf diese Weise wird eine zusätzliche Untersuchung zur Feinheit der definierten Aktivitäten ermöglicht.

Eine Herausforderung dabei, den so definierten Aktivitätskatalog innerhalb eines automatisierten Systems mit Computer Vision zu nutzen, ist die Erstellung eines gelabelten Datensatzes zum Ermöglichen des Trainings eines Netzwerks. Innerhalb des Projekts wurde sich dafür entschieden, einen Ansatz zur Analyse des Gesamtbildes der Bucht zu verfolgen.

Eine Alternative bietet der Einzeltieransatz, bei dem die einzelnen Tiere aus dem Gesamtbild ausgeschnitten werden und diesen Ausschnitten daraufhin eine oder mehrere Aktivitäten zugewiesen werden. Hierbei fällt die örtliche Zuweisung weg, da diese bereits in einem vorgelagerten Schritt durch das Ausschneiden des Tieres aus dem Originalbild geschieht.

Ein Vorteil der Gesamtbildanalyse besteht darin, dass hier für alle Schweine einer Bucht potenziell nur ein Netzwerk genutzt werden muss. Dieses kann den Rechenaufwand reduzieren. Bei einem Einzeltieransatz müssten in dem hier vorliegendem Versuchsfall bis zu 13 verschiedene Schweine pro Frame erkannt, ausgeschnitten und analysiert werden, was auch bei einem potenziell leichtgewichtigerem Analysenetzwerk einer Vervielfachung des Rechenaufwands entspricht. Da eine solche Analyse allerdings langfristig auf einem Edge-Gerät im Stall durchgeführt werden soll, wurde in dieser Arbeit darauf geachtet, die allgemeinen Rechen- und Hardwareanforderungen möglichst gering zu halten.

Zusätzlich ermöglicht die Nutzung des gesamten Bildes voraussichtlich bessere Rückschlüsse auf Aktionen mit mehreren Akteuren. Eine Vielzahl interessanter Schweineaktivitäten besitzen einen Aggressor und ein Opfer, bei denen das Zusammenspiel dieser beiden Tiere besonders interessant ist. Bei einem Einzeltieransatz könnten so bereits vor der Analyse durch ein neuronales Netzwerk wichtige Informationen verworfen werden.

Wie bereits erwähnt, erfordert der hier gewählte Ansatz zur Analyse des Gesamtbildes eine ST-Annotation des Datensatzes. Hierzu wird in diesem Paper eine eventbasierte Keyframe-Annotation mit einem unterstützenden YOLO-Netzwerk genutzt. Im Detail bedeutet dies, dass die zugrunde liegende Daueraufnahme in einzelne, kurze Events unterteilt wird. Diese Events werden von Hand ausgewählt und beinhalten ein möglichst aktives Verhalten von Schweinen innerhalb der Bucht. Diese Auswahl wird so durchgeführt, um eine möglichst große Vielfalt an Aktionen innerhalb eines Events und auch übergreifend über alle ausgewählten Events zu erhalten.

Um die so ausgewählten Events innerhalb eines Deep-Learning Netzwerks nutzen zu können, muss jedem Schwein eine Bounding Box zugeordnet sein, welcher dann

wiederum eine Auswahl an Aktivitäten angeheftet sind. Zusätzlich müssen für das Training alle Events die gleiche Dauer besitzen. Um diese Uniformität der Länge zu erreichen, wird jedes Event in einzelne gleichlange Clips zerlegt, deren Dauer durch den Keyframeabstand festgelegt wird. Hierbei muss dieser Abstand möglichst groß gewählt werden, um überflüssigen Annotationsaufwand zu verhindern. Gleichzeitig dürfen die Abstände nicht zu groß gewählt werden, um mögliche Aktionswechsel einzelner Tiere innerhalb eines Events nicht zu verpassen. Hierzu wurde in einer vorausgehenden Masterarbeit ein optimaler Keyframe-Abstand von 90 Frames ermittelt.

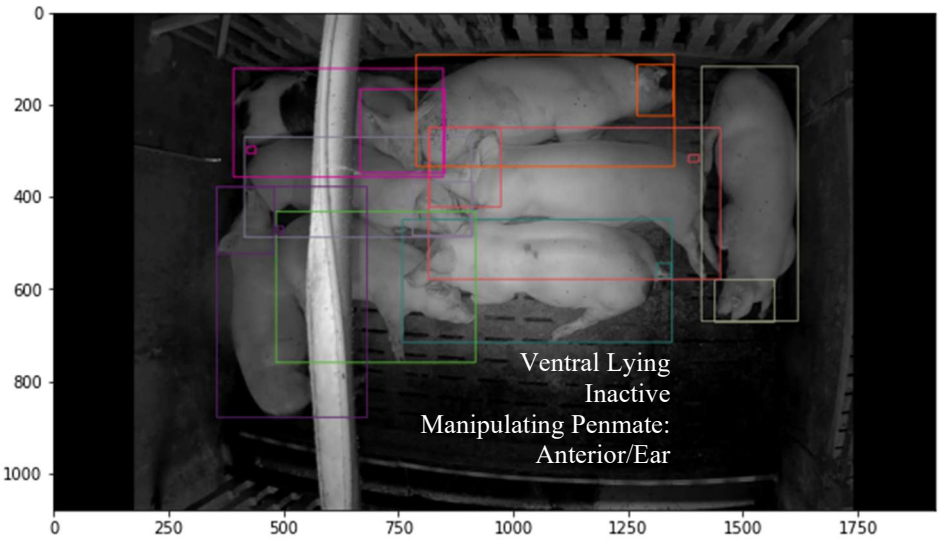

Abb. 1: Beispielbild der Daueraufnahme mit generierten Bounding Boxen und beispielhaft angefügten Aktivitätslabeln für die blaue Bounding Box

Um im eigentlichen Labeling-Prozess jeden Clip zu labeln, wird auf eine Kombination von Keyframes und einem trainierten YOLO-Netzwerk zur Schweineerkennung gesetzt. In einem ersten Schritt werden alle Frames der ausgewählten Events durch das YOLO-Netzwerk vorannotiert, wodurch für jeden Frame bereits Bounding Boxen der Schweine vorliegen. Im Folgenden können dann für jeden Keyframe die einzelnen Aktivitäten den ermittelten Bounding Boxen zugewiesen werden. Eine vereinfachte Darstellung eines so annotierten Frames mit Labeln für eine Box kann in Abb. 1 gesehen werden. Der hierdurch entstandene Datensatz besteht aus 349 Events und zugehörigen Keyframes, die eine Unterteilung in Clips ermöglichen. Alle Daten wurden im AVA-Format [Gu18] annotiert.

4.2 Genutzte Netzwerk-Architekturen

In diesem Abschnitt werden die in dem Projekt betrachteten Netzwerk-Architekturen vorgestellt, mit denen Untersuchungen zur Machbarkeit einer generellen Aktivitätserkennung durchgeführt wurden. Hierbei handelt es sich um drei spezielle

Implementierungen von state-of-the-art Ansätzen zur Aktivitätserkennung bei Menschen, welche von der Facebook Research Group erstellt wurden. Im Speziellen wird das X3D Netzwerk, beruhend auf 3D-Convolution, das SlowFast Netzwerk, basierend auf Two-Stream Architekturen und das MViT Netzwerk aus der Gruppe der Transformatoren genutzt.

Nachfolgend werden lediglich die grundlegenden Funktionsweisen sowie Vor- und Nachteile der Architekturen aufgezeigt. Generell konnte ausgehend von der existierenden Literatur nicht darauf geschlossen werden, welches Netzwerk die besten Ergebnisse erzielen könnte, da alle Benchmarking-Messwerte von menschlichen Aktionen ausgehen und Tiere generell außer Acht gelassen wurden.

Die sogenannten 3D-Convolution Networks (3D-CNN) bilden einen direkten Übergang der klassischen Bildanalysemethoden, wie CNNs, hin zur Videoanalyse. Hierbei wird ein Video als eine Reihe an Bildern aufgefasst, die neben den räumlichen Dimensionen Höhe und Breite auch eine temporale Dimension besitzen [Tr15]. Hierzu müssen die typischen Filter eines CNNs um eine dritte, zeitliche Dimension erweitert werden, sodass dreidimensionale Bildfilter entstehen. Diese Erweiterung bringt einige Vorteile mit sich.

So ist es möglich, bereits bekannte, erprobte Netzwerk-Architekturen aus der Bildanalyse direkt in den 3-dimensionalen Raum zu überführen. Hierbei können zusätzlich bereits vortrainierte Gewichte der einzelnen Schichten erweitert und übernommen werden [CaZi17]. Da das Feld der Videoanalyse mit neuronalen Netzwerken noch nicht so weit fortgeschritten ist wie die Bildanalyse, lohnt es sich oft, diese Gewichte zu übernehmen.

Ein Problem dieser Architekturen ist allerdings die Lokalität der Filter. Da die genutzten Filter nur eine begrenzte räumliche Dimension abdecken können, sind 3D-CNNs nicht gut in der Lage, zeitlich oder räumlich voneinander entfernte Merkmale in einen Zusammenhang zu bringen. Ob dieses Merkmal auch im Bereich der Schweineforschung zum Tragen kommt, ist jedoch fraglich, da sich die Tiere räumlich oft wenig bewegen und die Aktionen zeitlich direkt zusammenhängen. Das hier genutzte X3D Netzwerk [Fe20] ist eine optimierte Form eines 3D-CNNs, welche sich während des Trainingsprozesses den genutzten Trainingsdaten anpasst, um so ein möglichst effizientes Netzwerk zu bilden.

Neben 3D-CNNs gibt es in der Literatur einen weiteren vorherrschenden Ansatz zur Verarbeitung der zeitlichen Dimension in einem Video, die sogenannten Two-Stream-Architekturen. Bei diesem Modell werden zwei Netzwerke parallel genutzt, um jeweils den räumlichen Zusammenhang innerhalb eines Frames und den temporalen Zusammenhang mehrerer Frames zu ermitteln.

Der sogenannte „Spatial Stream“ oder auch „Slow Pathway“ einer solchen Architektur arbeitet auf individuellen oder wenigen Frames eines Videos und führt eine Aktivitätserkennung aus. Es hat sich dabei gezeigt, dass bereits statische Bilder Rückschlüsse auf Aktivitäten erlauben und sich hier die Nutzung eines 3D-CNNs als Pathway lohnt.

Dem gegenüber liegt der „Temporal Stream“ oder auch „Fast Pathway“. Ziel dieses Pfades ist es, eine möglichst genaue Repräsentation der temporalen Dimension zu verarbeiten, was durch eine möglichst hohe Bildrate erreicht wird. Auch hier wird wieder ein 3D-CNN als Pathway genutzt. Um bei dieser höheren Framerate die Rechenanforderungen vergleichbar mit dem „Spatial Stream“ zu halten, werden innerhalb dieses Pfades nur wenige Filter innerhalb einer Schicht verwendet.

Diese Struktur der zwei Pfade bringt einige Vorteile mit sich. Zum einen wird durch den dedizierten Pfad ein besseres temporales Verständnis erreicht, zum anderen können für beide Pfade vergleichbar einfache 3D-CNN Architekturen genutzt werden, da diese nur eine Art von Merkmalen erkennen sollen. Das hier genutzte SlowFast Netzwerk [Fe19] zeichnet sich zudem über einige Querverbindungen aus, durch die temporaler Kontext an die Aktivitätserkennung weitergegeben werden kann, schon bevor die beiden Pfade endgültig zusammengeführt werden.

Einen grundsätzlich anderen Ansatz verfolgen die Visual Transformer Networks (ViT) [Do20]. Diese basieren auf dem Prinzip der Self-Attention und kommen ursprünglich aus dem Bereich des Natural Language Processing. Hierbei wird die ursprüngliche Encoder-Decoder-Architektur an die neue Aufgabe der Bilderkennung angepasst, indem zum einen der Decoder entfernt und durch eine Klassifikation ersetzt wird. Zum anderen muss der Input angepasst werden.

Da die vorliegende Architektur eine Sequenz an Wörtern erwartet, muss die Bildsequenz in ein passendes Format gebracht werden. Hierzu werden einzelne Bilder fragmentiert und begradigt, wodurch die einzelnen Farbwerte eines Bildausschnitts als Token-Input des Netzwerks genutzt werden können.

Generell bietet der Ansatz der ViT eine Möglichkeit zur Erkennung besonders weit auseinanderliegender Zusammenhänge, egal ob örtlich oder zeitlich. Das hier genutzte Netzwerk der Multiscale Visual Transformer (MViT) [Fa21] bietet eine Architektur mit mehreren verschiedenen Skalierungsebenen, wodurch kleine bildliche Merkmale nach und nach in einen größeren temporalen Zusammenhang gebracht werden können.

5 Ergebnisse und Diskussionen

Vor einer Beurteilung der einzelnen Netzwerkgenauigkeiten muss der zugrunde liegende Datensatz dargestellt und mögliche Schwachpunkte müssen erläutert werden. Tab. 1 stellt eine vereinfachte Auflistung der Aktivitäten mit ihren jeweiligen Häufigkeiten dar. Hierbei ist zu erwähnen, dass nicht alle elementaren Aktivitäten dargestellt werden konnten. So bilden die Punkte „Manipulating“, „Manipulated“, „Performed Aggressive Behaviour“ und „Received Aggressive Behaviour“ weitere Oberkategorien, zu denen es einzelne elementare Aktivitäten gibt. Für diese Kategorien wurden die Häufigkeiten der darunterliegenden Aktivitäten akkumuliert. So stellt diese Tabelle die zweite von drei möglichen Ebenen auf, in die dieser Datensatz unterteilt werden kann. Elementare Aktivitäten, für die keine Events im Videomaterial vorhanden sind, wurden nicht berücksichtigt.

Aktivität	**#Events**	**#Keyframes**	**#Label**
Posture	--	--	--
Lateral Lying	346	2112	11444
Ventral Lying	240	1623	6144
Sitting/Kneeling	102	676	2141
Standing	316	1772	6857
Performed Behavioural States	*793*	*5630*	*20526*
Inactive	287	2075	14364
Locomotion	75	369	2019
Playing	26	201	267
Mounting	6	39	39
Manipulating	*399*	*2946*	*3837*
Received Behavioural States	*261*	*1908*	*2265*
Mounted	6	39	39
Manipulated	*255*	*1869*	*2226*
Performed Behavioural Events	*162*	*990*	*990*
Tail Biting	23	118	118
Ear Biting	42	260	260
Performed Aggressive Behaviour	*97*	*612*	*612*
Received Behavioural Events	*162*	*990*	*1046*
Tail Bitten	23	118	118
Ear Bitten	42	260	260
Received Aggressive Behaviour	*97*	*612*	*668*

Tab. 1: Übersicht des Aktivitätskatalogs mit zugehörigen Events, Keyframes und Labeln

Ausgehend von diesem Datensatz wurden die drei verschiedenen Modellarchitekturen angelernt und hingehend ihrer Genauigkeit verglichen. Hierzu wurden für alle Modelle vortrainierte Gewichte des AVA-Datensatzes [Gu18] genutzt und alle Modelle wurden auf allen drei möglichen Verallgemeinerungsstufen des Aktivitätenkatalogs getestet.

Modell	**Trainings-mAP**	**Test-mAP**
MViT	0.1534	0.1198
SlowFast	0.1294	0.1043
X3D	0.0953	0.0523

Tab. 2: Genauigkeiten der verschiedenen Netzwerke für alle möglichen Aktivitäten

In einem ersten Versuch wurden die Netzwerkarchitekturen auf allen zur Verfügung stehenden Aktivitätsklassen getestet, um herauszufinden, ob diese bereits bei der feinsten Aktivitätsgliederung hinreichend genaue Ergebnisse liefern. Hierbei wurde als Basis zur Bestimmung der Genauigkeit die mean Avergage Precision (mAP) gewählt. Anhand von Tab. 2 kann erkannt werden, dass alle Netzwerke ähnliche Genauigkeiten erreichen, wobei MViT mit 11,98% Testgenauigkeit am besten funktioniert. Allerdings muss gesagt werden, dass diese Genauigkeit nicht als akzeptabel gewertet werden kann, obwohl die erreichte Genauigkeit bei 30 Aktivitäten für ein kontextbasiertes Lernen spricht.

Modell	**Trainings-mAP**	**Test-mAP**
MViT	0.6320	0.5270
SlowFast	0.5821	0.5194
X3D	0.5443	0.5054

Tab. 3: Genauigkeiten der verschiedenen Netzwerke für die acht Oberkategorien

In einem weiteren Versuch wurden die maximal zusammengefassten Aktivitäten getestet, bei denen lediglich die Körperhaltungen und restlichen Oberkategorien genutzt werden. Hier zeigt sich eine klare Steigerung der Genauigkeit bei allen Netzwerken, wobei MViT auch hier am besten abschneidet, wie in Tabelle 3 gesehen werden kann.

Der Grund für diesen Genauigkeitsverlauf zwischen den verschieden genauen Aktivitätskatalogen lässt sich erkennen, wenn die erreichten Genauigkeiten pro Aktivität betrachtet werden. Hier zeigt sich, dass außer den Körperhaltungen keine elementare Aktivität mit mehr als 6 % Genauigkeit erkannt werden konnte. Dies bedeutet, dass die Netzwerke für diese keine kontextbasierten Merkmale lernen konnten und lediglich zufällig entscheiden. So lässt sich auch der Genauigkeitssprung bei einer Reduzierung auf acht Klassen erklären, da hier die hohen Genauigkeiten der Körperhaltungen die schlechten Genauigkeiten der restlichen Klassen ausgleichen.

Für dieses Verhalten kann es zwei mögliche Folgerungen geben. Zum einen ist es möglich, dass der hier verfolgte Ansatz der Gesamtbilderkennung nicht zielführend ist. Zum anderen kann es sein, dass durch die spärlich vertretenen Aktivitäten im Datensatz kein effektives Training ermöglicht wurde und der Datensatz zunächst ausbalanciert werden muss.

6 Schlussfolgerungen und Ausblick

Innerhalb dieses Papers konnten neben einem effektiven Weg zur Erstellung eines ST-annotierten Datensatzes für Schweine mehrere trainierte Netzwerke miteinander verglichen werden. So konnte eine automatisierte Erkennung von Aktivitäten bis zu einem gewissen Grad erreicht werden. Jedoch entsprechen diese Ergebnisse noch nicht der im Projekt angestrebten Genauigkeit. Die Verallgemeinerung einzelner Aktivitäten in allgemeine Oberkategorien verbesserte die Ergebnisse leicht, wobei jedoch nur für verschiedene Körperhaltungen eine echte Erkennung erreicht werden konnte. Aus diesem

Grund war die Ermittlung eines am besten geeigneten Netzwerks nicht möglich, weshalb alternative Ansätze verfolgt werden sollten.

Aktuell werden im Projekt Untersuchungen zur Nutzung von Positions-Heatmaps und Ellipsenrepräsentationen von Einzeltieren durchgeführt. Hierbei wird langfristig versucht, über diese Heatmaps Rückschlüsse auf den generellen Aktivitätslevel eines Tieres zu schließen.

Ein anderer Ansatz ist die Erkennung von paarweisen Interaktionen. Aggressive Verhaltensweisen unter Schweinen besitzen sowohl ein Opfer als auch einen Täter. Daher liegt der Gedanke nahe, diese paarweisen Interaktionen genauer zu untersuchen, um eine Erkennung von aggressivem Verhalten zu ermöglichen.

Fördernhinweis: Wir danken der Europäischen Innovationspartnerschaft „Produktivität und Nachhaltigkeit in der Landwirtschaft“ (EIP Agri) für die Förderung des Projektes SmartTail, im Zuge dessen diese wissenschaftliche Veröffentlichung entstehen konnte.

Literaturverzeichnis

[AKB20] Alameer, A., Kyriazakis, I., & Bacardit, J.: Automated recognition of postures and drinking behaviour for the detection of compromised health in pigs. Scientific reports, 10(1), S. 1-15, 2020.

[AkKu18] Aktionsplan Kupierverzicht, https://www.ringelschwanz.info/weitere-infomationen/-aktionsplan-kupierverzicht.html, Strand: 28.10.2022

[CaZi17] Carreira, J.; Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, S. 6299-6308, 2017.

[Ch20] Chen, C. et al.: Recognition of feeding behaviour of pigs and determination of feeding time of each pig by a video-based deep learning method. *Computers and Electronics in Agriculture*, *176*, 105642, 2020.

[CZN21] Chen, C.; Zhu, W.; Norton, T.: Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning. Computers and Electronics in Agriculture 187, 2021.

[Do20] Dosovitskiy, A. et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[Fa21] Fan, H. et al.: Multiscale vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, S. 6824-6835, 2021.

[Fe19] Feichtenhofer, C. et al.: Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision, S. 6202-6211, 2019.

[Fe20] Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, S. 203-213, 2020.

[Gu18] Gu, C. et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, S. 6047-6056, 2018.

[Hu20] Huang, J. et al.: Automated video behavior recognition of pigs using two-stream convolutional networks. *Sensors*, *20*(4), S. 1085, 2020.

[Li20] Liu, D. et al.: A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs. Biosystems Engineering, 195, S. 27-41, 2020.

[Na19] Nasirahmadi, A. et al.: Deep learning and machine vision approaches for posture detection of individual pigs. Sensors, 19(17), S. 3738, 2019.

[SiZi14] Simonyan, K.; Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27, 2014.

[Tr15] Tran, D. et al.: Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, S. 4489-4497, 2015.

[Ya18] Yang, A. et al.: Automatic recognition of sow nursing behaviour using deep learning-based segmentation and spatial and temporal features. Biosystems Engineering, 175, S. 133-145, 2018.

[Ya20] Yang, A. et al.: An automatic recognition framework for sow daily behaviours based on motion and image analyses. Biosystems Engineering, 192, S. 56-71, 2020.

[Zh18] Zheng, C. et al.: Automatic recognition of lactating sow postures from depth images by deep learning detector. Computers and electronics in agriculture, 147, S. 51-63, 2018.

[Zh20] Zhang, K. et al.: A spatiotemporal convolutional network for multi-behavior recognition of pigs. *Sensors*, *20*(8), S. 2381, 2020.

[Zo11] Zonderland, J. J et al.: Characteristics of biter and victim piglets apparent before a tail-biting outbreak. Animal, 5(5), S. 767-775, 2011.