Schlussbericht zur Machbarkeitsstudie zur Nutzung KI-basierter Sensorik zur Tunneleingangsüberwachung TuNuKi

11. Dezember 2023

1 Aufgabenstellung

Ziel der Machbarkeitsstudie war die Untersuchung der Frage, wie eine automatische Tunneleingangsüberwachung mithilfe von KI durchgeführt werden kann. Die Daten, auf welchen die KI arbeitet, stammen zum einen von regulären Überwachungskameras und zum anderen von Event-Kameras, auch Dynamic Vision Sensor(DVS) genannt, welche anstelle von vollständigen Bildern mit einer festen Bildrate nur Änderungen in der Szene mit extrem hoher zeitlicher Auflösung ausgeben. Für beide Sensortypen wurden Algorithmen entwickelt, welche die Präsenz von Personen am Bahntunneleingang feststellen können. Dabei bestand das Hauptziel darin, die beim aktuell im Einsatz befindlichen System hohe Fehlalarmrate zu reduzieren.

2 Voraussetzungen des Vorhabens

Am Testtunnel existiert bereits ein System zur automatischen Erkennung von unbefugtem Eindringen, welches auf klassischen Überwachungskameras und LiDAR-Sensoren(Light detection and ranging) basiert. Dieses System löst allerdings häufig, laut Schätzung der Bundespolizei ungefähr 10 mal im Monat, Fehlalarme aus. Die Ursache liegt meist bei widrigen Wettereinflüssen oder Tieren am Tunneleingang. Die Ungenauigkeit ist unter anderem damit zu begründen, dass für die korrekte Einordnung von vorbeifahrenden Zügen eine hohe zeitliche Auflösung notwendig ist, welche beim Einatz von LiDAR zu Lasten der räumlichen Auflösung erreicht wird. Damit werden kleinere Objekte, wie Personen, nur noch durch wenige Scanpunkte in der aufgenommenen Punktwolke dargestellt.

Der in der Machbarkeitsstudie verwendete Dynamic Vision Sensor (Metavision® EVK3 – Gen4.1) ist kommerziell verfügbar. Ein DVS nimmt nicht wie eine herkömmliche Frame-Kamera in einem festen Intervall vollständige Bilder auf. Stattdessen werden durch die einzelnen Pixel assynchron und unabhängig Events ausgegeben, sobald in dem entsprechenden Pixel eine Helligkeitsänderung

auftritt. Das führt dazu, dass nur Änderungen in der Szene aufgezeichnet werden. Zusätzlich wir dadurch die Aufzeichnung der Daten mit einer extrem hohen zeitlichen Auflösung ermöglicht, da nicht an einer festen Bildrate festgehalten wird. Der Sensor hat ein Pixelarray mit 1280×720 Pixeln. Im Rahmen der Machbarkeitsstudie wurde das Metavision SDK [Prophesee, 2023] des selben Herstellers eingesetzt, um die Aufnahmesoftware zu implementieren und die Eventdaten aufzubereiten.

Bei der Erfassung von traditionellen Videos im Rahmen des Projektes wurde die bestehende Infrastruktur ausgenutzt. Es wurde festgestellt, dass die Deutsche Bahn bereits an strategischen Punkten, besonders an den Ein- und Ausgängen der Bahntunneln, Framekameras installiert hatte. Anstatt also eigene Kameras aufzubauen, wurde der pragmatische Ansatz gewählt, auf diese bestehenden Kamerasysteme zurückzugreifen. Dieser Ansatz bot nicht nur finanzielle Vorteile durch die Reduzierung der Projektkosten, sondern ermöglichte auch eine rasche Umsetzung, da keine neuen Installationen erforderlich waren.

3 Projektplanung und Ablauf

Der Projektantrag wurde Anfang August 2022 genehmigt. Bereits im August 2022 fanden mehrere Meetings mit der Deutschen Bahn und der Bundespolizei statt, um einen passenden Eisenbahntunnel für die Durchführung der praktischen Untersuchungen zu finden. Nach drei Meetings fiel die Wahl auf den Tunnel am Berliner Flughafen. Um sicherzustellen, dass dieser Tunnel für das Vorhaben geeignet ist, trafen sich alle am Projekt beteiligten Partner am 30. August 2022 vor Ort zu einer Begehung. Im Rahmen dieser Begehung wurden erste Testmessungen durchgeführt und die möglichen Lösungen für das Anbringen des DVS und das Aufstellen der Computertechnik diskutiert. Zusätzlich wurden von der Bundespolizei nachts weitere Testmessungen durchgeführt, um den Einfluss der aktuell genutzten LiDAR-Systeme auf die DVS-Daten bei Dunkelheit zu untersuchen. Nach Auswertung der Testmessungen wurde der ausgewählte Tunnel als geeignet bewertet. Danach erfolgte von Seiten der Deutschen Bahn die Beschaffung der notwendigen Genehmigungen und die Organisation des Installationstermins, da zum Anbringen des DVS aus Sicherheitsgründen eine Streckensperrung erforderlich war. Dieser Termin wurde dann auch für das Aufnehmen von ersten Szenen mit Personen im Kamerasichtfeld genutzt.

Da sich bei der Auswertung der Aufnahmen herausstellte, dass der ausgewählte Tunnel nur sehr selten von Wartungspersonal betreten wird, wurde ein zusätzlicher Termin zum Aufnehmen von Trainingsdaten mit Personen im Tunnel von Seiten der Bundespolizei und der Deutschen Bahn organisiert, für den wiederum eine Streckensperrung erforderlich war.

Im Rahmen des Projekts wurden drei große Meetings (Auftakt-, Zwischen- und Abschlussmeeting) am 24. Oktober 2022, am 5. Juli 2023 und am 6. Oktober 2023 durchgeführt. Daneben gab es zahlreiche Meetings mit den assoziierten Partnern bis zum Anbringen der Kamera am 21. Februar 2023. Die beiden Projektpartner selbst haben sich regelmäßig einmal monatlich bezüglich des

Projektfortschritts abgesprochen.

Die Datenübermittlung an die Projektpartner wurde von der Bundespolizei übernommen. Alle zwei Wochen wurde während der Aufzeichnung der Testdaten mit dem DVS zwischen Februar und Mai 2023 die externe Festplatte am Computer im BER-Tunnel gewechselt und nach Krefeld geschickt. Der Austausch der Daten von der Überwachungskamera erfolgte einmalig über eine Blu-Ray Disc.

4 Stand der Technik

4.1 Ausgangslage bei der Tunneleingangsüberwachung in Hinblick auf die verwendete Sensorik

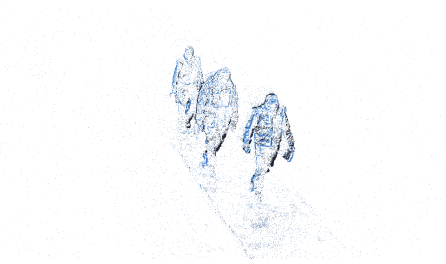
Die im Projekt behandelte Thematik der Unterscheidung zwischen Personen und anderen sich bewegenden Objekten wie Zügen und Tieren an Tunneleingängen ist nicht einfach, da dort meist starke Witterungseinflüsse und ungünstige Lichtverhältnisse herrschen. Insbesondere die starke Überbelichtung von CCTV-Kameras aufgrund der großen Helligkeitsunterschiede am Tunneleingang ist ein Problem bei der Verwendung herkömmlicher Bildverarbeitungsalgorithmen. Weiterhin werden auch PTZ-Kameras mit integrierten Bewegungstrackern eingesetzt. Die Probleme sind jedoch ähnlich denen der CCTV-Kameras.

Daher werden zunehmend LiDAR-Sensoren (Light detection and ranging) genutzt, die unabhängig von der Umgebungshelligkeit arbeiten. Dabei werden Laserstrahlen aktiv ausgesendet und aus der Lichtlaufzeit der Signale die Entfernungen der Objekte berechnet. LiDAR-Sensoren führen somit ein dreidimensionales Laserscanning durch. Problematisch bei dieser Technologie ist jedoch die Abhängigkeit zwischen erreichbarer räumlicher und zeitlicher Auflösung. Bei einer hohen zeitlichen Auflösung, wie sie für die Identifikation schnell fahrender Züge notwendig ist, ist nur eine geringe räumliche Auflösung realisierbar, sodass kleinere Objekte, wie z.B. weiter entfernte Personen, nur wenige Scanpunkte in der aufgenommenen LiDAR-Punktwolke besitzen. Eine sichere Detektion von sich bewegenden Personen ist somit nicht möglich, was zu einer hohen Anzahl von Fehlalarmen, aber auch von Nichtdetektionen führt.

Eine weitere Möglichkeit zur Überwachung von Tunneleingängen besteht in der Nutzung von Infrarot-Schranken [Setola et al., 2015]. Hier kann jedoch nicht zwischen Person und durchfahrendem Zug unterschieden werden. Außerdem können so nur Eintritte erkannt werden, die nicht gleichzeitig mit Zugverkehr erfolgen.

Zum Umgehen dieser Problematik könnten Sensorfußmatten [Catalano et al., 2017] genutzt werden, da diese nur den begehbaren Bereich überwachen. Einerseits stellt hier die Klassifizierung des auslösenden Signals (Unterscheidung des Betretens der Matte z.B. durch eine Person oder ein Tier) eine größere Herausforderung dar und andererseits könnten sich Personen im Gleisbett bewegen, wenn gerade kein Zugverkehr erfolgt.

(a) Beispielaufnahme von einem DVS in Bewegung



(b) Beispielaufnahme von einem stationärem DVS

Abbildung 1: Vergleich von Aufnahmen durch stationäre und bewegte DVS

4.2 Ausgangslage bezüglich Nutzung von Dynamic Vision Sensoren

Ein vorherrschender Beweggrund für den Einsatz des Dynamic Vision Sensors in der Personenerkennung findet sich im Automobilkontext. Prophesee, der Hersteller des in dieser Studie verwendeten DVS, stellt zwei große Datensätze in diesem Bereich zur Verfügung. Der eine ist der „GEN1 Automotive Detection“ Datensatz [de Tournemire et al., 2020] mit manuell erstellten Bounding Box Labels. Der andere ist der höher aufgelöste „1 Megapixel Automotive Detection“ Datensatz [Perot et al., 2020] mit Bounding-Box Labels, die mittels eines etablierten KI-Verfahrens aus einer parallel mit einer Frame-Kamera aufgenommenen RGB-Aufnahme extrahiert wurden. Die meisten Daten in diesen Datensätzen wurden mit einem bewegten DVS aufgenommen. Dies führt dazu, dass unbewegliche Objekte in der Umgebung viele Ereignisse erzeugen. Diese Eigenschaft führt zu einem erheblichen Unterschied zwischen den Daten in diesen Datensätzen und unserem Anwendungsfall, wie in Abbildung 1 durch den Vergleich eines Event-Frames aus dem „1 Megapixel Automotive Detection“ Datensatz und unserem Datensatz gezeigt wird. Darüber hinaus führt die Positionierung des DVS dazu, dass Personen auf gleicher Höhe mit dem DVS aufgezeichnet werden, während Überwachungskameras normalerweise eine Ansicht von oben herab aufzeichnen. Die Anwendung des RED-Modells [Perot et al., 2020] für Bounding-Box-Erkennung, das auf dem „1 Megapixel Automotive Detection“ Datensatz trainiert wurde, auf unsere Daten hat gezeigt, dass die auf dem ersten Datensatz gelernten Merkmale nicht gut auf unseren Anwendungsfall übertragbar sind.

[Jiang et al., 2019] untersucht die Kombination von aus DVS-Signalen berechneten Konfidenzwerten pro Pixel und traditionellen Frames für die Bounding-Box Personenerkennung. Dies ist für uns nicht möglich, da wir nicht in der Lage waren, frame-basierte Daten aufzuzeichnen, welche eine einfache räumliche Zuordnung zulassen. [Miao et al., 2019] bietet einen kleinen 346×260 Benchmark-Datensatz mit 12 Clips von ca. 30 Sekunden für die Bounding-Box-Personenerkennung in verschiedenen Einstellungen. [Bisulco et al., 2020] untersucht die Bounding-Box Personenerkennung auf einem kleinen nicht-öffentlichen 480×320-Datensatz mit

Fokus auf Bandbreitenreduktion. [Wan et al., 2021] stellt einen 488 Sekunden langen Bounding-Box-Datensatz zur Personenerkennung mit einer Auflösung von 346×260 Pixeln zur Verfügung, der in verschiedenen Einstellungen aufgenommen wurde, und untersucht die Personenerkennung auf diesem Datensatz. [Alonso and Murillo, 2019, Bolten et al., 2021] stellen Datensätze mit halbautomatisch generierten semantischen Segmentierungslabels mit Fußgängern bereit. [Bolten et al., 2023] bietet einen Fußgängerdatensatz mit Instanzsegmentierungs-Labels, die durch die Aufnahme von Personen in leicht zu unterscheidenden Anzügen generiert wurden. Die geringeren Auflösungen und unterschiedlichen Szenarien, in denen diese Datensätze aufgenommen wurden, erschweren ihre Anwendung auf diesen Anwendungsfall.

In einem Überwachungskontext wird in [Perez-Cutino et al., 2021] die Erkennung von menschlichem Eindringen mit Hilfe von Dynamic Vision Sensors untersucht, wobei Daten mit einem auf einer Drohne montierten DVS erfasst und Personen in zwei Stufen erkannt werden. Zunächst werden bewegte Objekte erkannt und anschließend wird bestimmt, ob es sich um Menschen handelt.

4.3 Ausgangslage bezüglich Nutzung von Frame-Kameras

Bei der Entwicklung des Überwachungssystems griffen wir auf den neuesten Stand der Technik in der Objekterkennung zurück, insbesondere den YOLO (You Only Look Once) Algorithmus in seiner achten Version, YOLOv8. Das YOLO-Framework hat sich seit seiner Erstveröffentlichung als eines der führenden Systeme in der Echtzeit-Objekterkennung etabliert. Die achte Version, YOLOv8, bringt gegenüber seinen Vorgängern zahlreiche Optimierungen in Bezug auf Genauigkeit und Geschwindigkeit. Die Anwendung nutzte die öffentlich zugänglichen Gewichtungen und Konfigurationen des YOLOv8-Modells, die unter den Bedingungen der GNU General Public License (GPL) veröffentlicht wurden. Es ist wichtig zu betonen, dass, während wir auf das Basis-Modell von YOLOv8 zugriffen, alle Anpassungen und Optimierungen für unsere spezifischen Anforderungen - die Erkennung von Menschen und Zügen - eigens von uns durchgeführt wurden. Hierzu wurde ein Datenset erstellt und dem KI-Modell bereitgestellt, sodass auf die spezifisch verbauten Kameras innerhalb der Testanlage eingegangen werden konnte.

5 Zusammenarbeit mit anderen Stellen

Die Hochschule Niederrhein hat im Rahmen der Machbarkeitsstudie mit zwei assoziierten Projektpartnern zusammmengearbeitet, der DB Station&Service AG und der Bundespolizei.

In Zusammenarbeit mit der DB Station&Service wurde die Montage des DVS an einem von derselben betriebenen Bahntunneleingang geplant und umgesetzt, um Trainings- und Testdaten für das Projekt sammeln zu können. Nach einigen terminlichen Verzögerungen, da für die Montage des Sensors eine zeitweise Sperrung des Tunnels geplant werden musste, um diese sicher durchführen zu

können, wurde der Sensor in der Nacht vom 21.02.2023 zum 22.02.2023 erfolgreich, mit einer Verspätung von etwa 3 Monaten gegenüber der ursprünglichen Projektplanung¸ montiert.

In Zusammenarbeit mit der Bundespolizei wurde der Austausch der von dem DVS gesammelten Daten realisiert. Die Daten wurden von dem Aufnahmesystem auf eine externe USB-Festplatte geschrieben welche in einem zweiwöchigen Rhythmus von der Bundespolizei ausgetauscht und per post an die Hochschule Niederrhein gesendet wurde. Der Austausch der DVS Daten verlief ohne nennenswerte Schwierigkeiten.

Beide assoziierten Partner beteiligten sich an der Anforderungsspezifikation für das System. Dies betrifft vor allem die technische Realisierbarkeit und die von dem System zu bewältigenden Szenarien. Zusätzlich wurden durch Mitarbeiter der Bahn und der Bundespolizei einige von der Hochschule Niederrhein angefragten Szenarien am Tunneleingang gestellt, welche zur Bewertung des erstellten KI-Modells eingesetzt wurden.

Die von der Deutschen Bahn bereitgestellten Kameradaten stammten aus einem Netzwerk von Überwachungskameras, die strategisch an den Ein- und Ausgängen der Tunnel positioniert waren. Diese Kameras lieferten Video-Feeds in verschiedenen Auflösungen und Formaten. Die Daten wurden über die Software des Kameraherstellers Bosch in deren Videoformat bereitgestellt. Um eine effiziente Analyse zu ermöglichen, mussten wir uns mit dem Format, der Auflösung und den Übertragungsraten der Kameradaten auseinandersetzen. Die Vielfalt dieser technischen Spezifikationen erforderte eine sorgfältige Abstimmung, um die Kompatibilität mit unserem System sicherzustellen und die Leistungsfähigkeit des YOLOv8-Modells zu optimieren. Hierfür musste das Format zuerst in ein .avi Format konvertiert werden und konnte anschließend in die neuronalen Netze zur Verarbeitung übergeben werden. Der Umfang der Daten beschränkte sich auf die zuvor definierten Testszenarien, welche von Mitarbeitern der Deutsche Bahn sowie der Bundespolizei aufgenommen wurden.

6 Erzieltes Ergebnis

6.1 Verwendbarkeit des Dynamic Vision Sensors

Die Aufnahme der Daten am Tunneleingang erfolgte mit dem in Abbildung 2 dargestellten Aufbau. Die Sicht auf den Bordstein ist durch den Kabelkanal leider teilweise verdeckt. Dieses Problem sollte aber in Zukunft durch Modifikation der Anbringung gelöst werden. Der Versuch, durch den Kabelkanal größtenteils verdeckte Personen zu erkennen, würde die Fehlalarmrate in freier Sicht, welche bei der Planung eines Überwachungsaufbaus sichergestellt werden kann, unnötig erhöhen.

Die Anwendung neuronaler Netze auf kontinuierliche Event-Streams erfordert die Konvertierung der Ereignisse in ein Format mit einer festen Größe. Eine Möglichkeit, dies zu erreichen, wäre die Eingabe der Eventdaten als 3D-Punktwolken, wobei die räumlichen Koordinaten im Pixelarray (x, y) und der Zeitstempel t

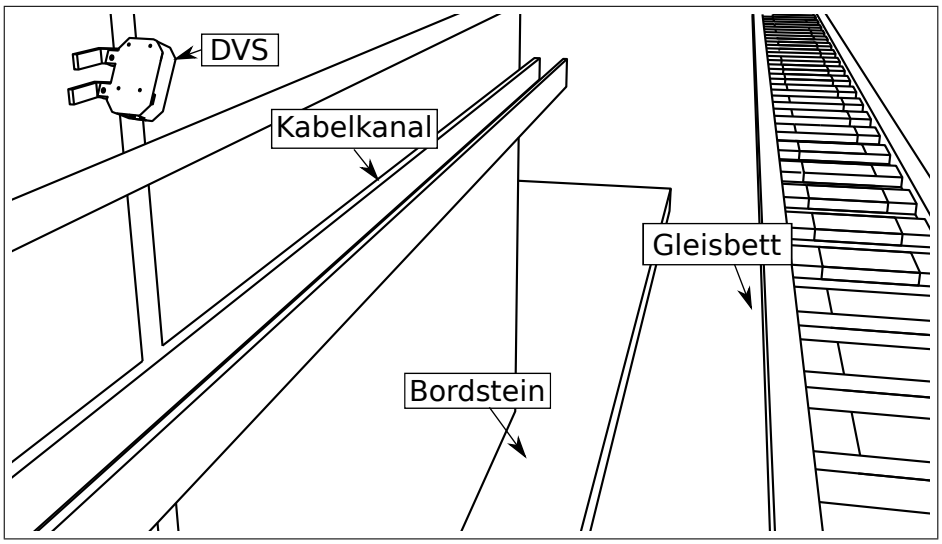


Abbildung 2: Anbringung des DVS am Tunneleingang.

die Punktkoordinaten bilden. Eine feste Eingabegröße wird dann erreicht indem Punkte mehrfach übergeben oder zufällig verworfen werden. Eine weitere Möglichkeit, welche weniger Rechenleistung erfordert, ist die Aufteilung des Ereignisstroms in zeitliche *Bins* einer bestimmten Länge und die Erzeugung einer zeitkodierten Darstellung aller Events in einem Bild. Wir wählten die Länge der Bins $T = 50000\mu s$, was zu 20 Bins pro Sekunde führte. Die in den Ereignissen enthaltenen räumlichen Informationen wurden verwendet, um ihre Position in der zeitkodierten Darstellung zu bestimmen. Es gibt verschiedene Optionen für den Einfluss des Zeitstempels t und der Polarität p auf die zeitkodierte Darstellung.

Linear Time Surface Die dichte Darstellung hat die Dimensionen $H \times W \times 2$, wobei H die Höhe und W die Breite des Pixelarrays ist, aus dem der Ereignisstrom stammt. Jeder Kanal ist einer der beiden möglichen Polaritäten zugeordnet. Der Wert an jeder Position wird entsprechend dem letzten Auftreten eines Ereignisses an dieser Position nach der folgenden Formel zugewiesen:

$$T(x, y, p) = \frac{t_{max(x,y,p)} - t_0}{T},$$

wobei $t_{max(x,y,p)}$ der Zeitstempel des letzten Ereignisses an der Position (x, y) mit Polarität p, t_0 der Zeitstempel zu Beginn des aktuellen Timebins und T die Länge eines Timebins ist. Während die Dokumentation des Metavision SDK des Sensorherstellers [Prophesee, 2023] diese Kodierung als *Linear Time Surface* bezeichnet[1], wird dieser An-

[1] https://docs.prophesee.ai/stable/tutorials/ml/data_processing/event_preprocessing.html

satz in der Literatur oft als *surface of active events (SAE)* bezeichnet [Wan et al., 2021, Mueggler et al., 2015, Benosman et al., 2014].

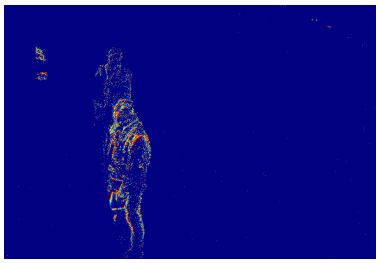
Event-Volume Ein Event-Volume [Zhu et al., 2019] repräsentiert die räumliche Position jedes Ereignisses in den ersten beiden Dimensionen und stellt den Zeitstempel als Kombination aus der dritten Dimension und dem Wert dar. Die dritte Dimension unterteilt den Timebin weiter in separate Mikrobins, die nach der Polarität des Ereignisses aufgeteilt sind, was bedeutet, dass die Struktur im Wesentlichen aus separaten Event-Volumes für jede Polarität besteht. Jedes Ereignis verteilt dann einen Beitrag von eins auf die beiden nächstgelegenen Mikrobins, so dass die genaue Verteilung der Eingangsereignisse bis auf einen Rundungsfehler rekonstruiert werden kann, wenn jeder Punkt nur von einem Ereignis beeinflusst wird. Wir erzeugen das Event-Volume mit insgesamt sechs Mikrobins, drei für jede Polarität.

Wir führten die Kodierung mit den Metavision-Implementierungen durch. Aufgrund der künstlichen Beleuchtung am Tunneleingang waren die Aufnahmen von dort deutlich stärker verrauscht, als die inszenierten Aufnahmen in natürlich beleuchteten Szenen, welche für das Training verwendet wurden. Die Menge des Rauschens schwankte auch in Abhängigkeit davon, welche Beleuchtung zu einem bestimmten Zeitpunkt eingeschaltet war. Um diesen Unterschied zu unterdrücken, filterten wir räumlich-zeitlich Ereignisse heraus, für die in der 8-Punkte-Nachbarschaft innerhalb der letzten 50 ms keine anderen Ereignisse aufgetreten waren.

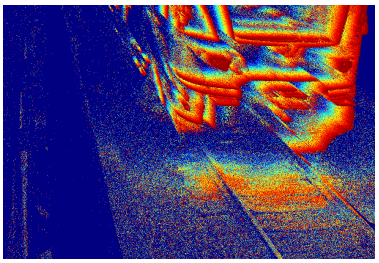
Eine Situation, die nicht direkt erfasst werden konnte, war das Betreten des Tunnels während der Durchfahrt eines Zuges. Auch wenn in dieser Situation genügend Platz vorhanden war, um auf dem Bordsteig in der Nähe der Tunnelwand zu gehen, wäre dies für einen Darsteller zu gefährlich. Aus diesem Grund waren wir nicht in der Lage, inszenierte Aufnahmen von dieser Situation zu erstellen. Bei dem Versuch, unbefugtes Betreten zu erkennen, konnte diese Situation jedoch nicht ausgeschlossen werden. Daher musste sie in den Datensatz aufgenommen werden. Dazu erzeugten wir die Beispiele künstlich, indem wir die erfassten Aufnahmen kombinierten.

Die Methode, die wir dazu verwendeten, war die Zusammenführung der Event-Streams vor der Kodierung in das Eingabeformat für die Detektoren. Da der Event-Stream, abgesehen von technischen Fehlern, immer nach t sortiert ist, wurde das Zusammenführen selbst wie ein Merge-Sort Schritt mit den Ereignisströmen als Eingabe durchgeführt, möglicherweise mit einem auf t angewendeten Offset, um Abschnitte zusammenzuführen, die zu unterschiedlichen Zeiten in ihren jeweiligen Aufzeichnungen auftreten. Beispielsweise ergibt die Kombination der in Abbildung 3a und Abbildung 3b dargestellten Ereignisströme Abbildung 3c.

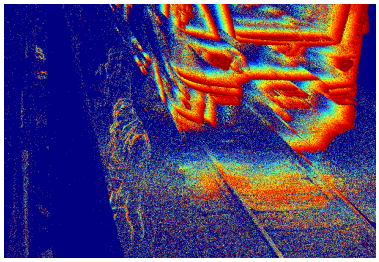
Bei dieser Methode wird die Verdeckung nicht berücksichtigt. Da Ereignisse, abgesehen von Rauschen, nur in nicht homogenen, sich bewegenden Bereichen erzeugt werden, erscheinen alle homogenen Teile transparent. Um hier Abhilfe zu schaffen, klassifizierten wir einen der zusammenzuführenden Event-Streams



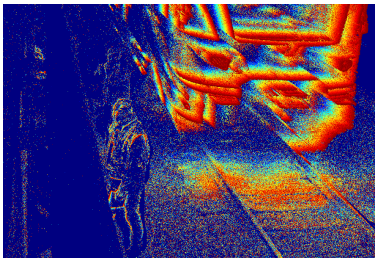
(a) Szene mit Personen.



(b) Szene mit vorbeifahrendem Zug.



(c) Zusammengeführte kodierte Streams ohne Verdeckung.



(d) Zusammengeführte kodierte Streams mit Verdeckung.

Abbildung 3: Beispiel für die Zusammenführung von Event Frames.

als *Vordergrund* und den anderen als *Hintergrund*. Nach der Rauschfilterung des Vordergrund-Ereignisstroms verwendeten wir eine morphologisches Closing, um eine Maske zu erhalten, welche die ungefähre Verdeckung durch sich bewegende *Vordergrund*-Objekte in der Szene zeigt. Ereignisse, die von der Maske abgedeckt wurden, wurden dann aus dem Hintergrundstrom entfernt oder beim Zusammenführen ignoriert. Das Ergebnis wird in Abbildung 3d dargestellt. Die Ereignisse, die hinter Personen auftreten, werden entfernt, aber einige Ereignisse werden fälschlicherweise an konkaven Abschnitten der Personenkonturen entfernt.

Es gibt noch einige Einschränkungen dieser Methode, die hier nicht gelöst wurden. Erstens erlaubt die Klassifizierung der einzelnen Ereignisströme als entweder „Vordergrund“oder „Hintergrund“ keine dreidimensionale Tiefe. Dies kann problematisch sein, wenn man z. B. versucht, Aufnahmen mit Regen oder Schnee zu erzeugen, bei denen man erwarten würde, dass einige Tropfen vor und andere Tropfen hinter die sich bewegenden Personen oder anderen Objekte fallen. Zweitens wirken sich größere Beleuchtungsänderungen in einer Szene oft nicht so auf die andere Szene aus, wie sie sollten. Wenn zum Beispiel ein Zug mit eingeschalteten Scheinwerfern vorbeifährt, könnte man erwarten, dass das vorbeifahrende Licht über Personen in der Nähe streift und Ereignisse erzeugt. Dieser Effekt wird durch das einfache Zusammenführen der Ereignisströme nicht reproduziert. Im Rahmen dieser Machbarkeitsstudie erachten wir dieses Vorgehen trotz dieser Einschränkungen als ausreichend, um den Lösungsansatz mit Hinblick auf diese sonst nicht aufzunehmende Situation zu bewerten.

Die Aufnahmen wurden in Clips unterteilt und manuell auf der Grundlage von Bildern, die aus dem Ereignisstrom generiert wurden, mit einem von zwei Labels versehen. Wir haben uns dafür entschieden, unser Problem als Bildklassifizierungsaufgabe und nicht als Objekterkennungsaufgabe zu formulieren, da wir nur an der Anwesenheit von Personen interessiert sind, nicht aber an deren genauer Position. Durch diese Vereinfachung des Problems konnten wir uns darauf konzentrieren, die Genauigkeit der Alarme zu verbessern. Die Auffassung des Problems als Segmentierungsaufgabe wurde ebenfalls untersucht. Allerdings wurde der Aufwand in der Datensatzerstellung als zu hoch eingeschätzt, um innerhalb der Projektlaufzeit verwertbare Ergebnisse zu erzielen. Das Label *People* wird Clips zugewiesen, in denen Personen vorkommen, während das Label *NoPeople* allen anderen Clips zugewiesen wird.

Bei der Aufteilung in einen Trainings- und einen Validierungsdatensatz ist es wichtig, für jeden Datensatz völlig unterschiedliche Clips auszuwählen. Eine zufällige Auswahl von generierten Frames aus den erhaltenen Aufnahmen würde die Ergebnisse verfälschen. Ein Detektor, der bei drei aufeinanderfolgenden Bildern overfitted ist, würde beispielsweise das mittlere Bild höchstwahrscheinlich immer noch richtig einstufen, wenn dieses im Training nicht gesehen wurde, die beiden anderen aber schon.

Der gelabelte Trainingsdatensatz enthält 20 Minuten und 15 Sekunden an nicht zusammengesetzten Aufnahmen, die als Personen enthaltend gekennzeichnet sind und auf dem Campus aufgenommen wurden, und 41 Minuten und 10 Sekunden an Aufnahmen, die als keine Personen enthaltend gekennzeichnet sind und sowohl auf dem Campus als auch am Tunneleingang aufgenommen wurden. Zusätzlich werden 1 Minute 42 Sekunden zusammengesetzte Aufnahmen generiert und sowohl als zusammengesetzte Bins als Beispiele mit Personen als auch als nur aus dem Hintergrund generierte Bins als Beispiele ohne Personen hinzugefügt. Um eine Überanpassung zu vermeiden, werden die Bins in einem Intervall ausgewählt, das sich an der Geschwindigkeit der Objekte in der Szene orientiert. Für die meisten Szenen, einschließlich derjenigen mit Menschen, wird jedes zwanzigste Bin ausgewählt, d. h. ein Bin pro Sekunde der Aufnahme. In Szenen mit Zügen, einschließlich zusammengesetzter Szenen, wurde jedes fünfte Bin ausgewählt. Für Szenen mit extrem kurzlebigen Ereignissen wie z. B. Lichtblitzen wurden sämtliche Bins übernommen.

Der Validierungsdatensatz enthält 9 Minuten und 16 Sekunden nicht zusammengesetzte Aufnahmen, die als Aufnahmen mit Personen gekennzeichnet sind und am Tunneleingang aufgenommen wurden und 2 Minuten und 48 Sekunden normale Aufnahmen, die als Aufnahmen ohne Personen gekennzeichnet sind. Wie für den Trainingsdatensatz wurden 27 Sekunden zusammengesetzter Aufnahmen als entsprechende Positiv- und Negativbeispiele hinzugefügt. Alle Beispiele im Validierungsdatensatz wurden am Tunneleingang aufgenommen. Auf die Validierungsdaten wurde keine Intervallabtastung angewendet.

Wir verwendeten drei verschiedene neuronale Netzwerkarchitekturen, um die zeitkodierten Repräsentationen des Event-Streams entweder als Personen enthaltend oder nicht Personen enthaltend zu klassifizieren.

Als Basisansatz mit geringer Komplexität betrachteten wir die folgende

Struktur:

- Ein Batch Normalization Layer
- Drei Encoder Blöcke aus:
  - Einem 2D CNN-Layer mit einem 3×3 Kernel
  - Einem 50% Dropout Layer
  - Einem 2×2 Max Pooling Layer
- Ein Dense Layer aus 10 Knoten
- Ein Outputknoten

Das Netz empfängt eine Eingabe der Form $360 \times 640 \times 2$ bei Time Surfaces oder der Form $360 \times 640 \times 6$ bei Event Volumes.

Wir untersuchten weiterhin die Leistung der MobileNetV2 [Sandler et al., 2018] Architektur. Da sie mit Hinblick auf Speichereffizienz entwickelt wurde, ist die Untersuchung ihrer Leistung nützlich, um eine effiziente endgültige Implementierung zu finden, die angesichts der Vollzeitüberwachungsanwendung wirtschaftlich relevant ist. Die Gewichte des Kodiererteils wurden mit den von TensorFlow [Russakovsky et al., 2015] zur Verfügung gestellten ImageNet-Gewichten initialisiert [Abadi et al., 2016]. Der Klassifikationskopf wurde durch einen Zwei-Klassen-Kopf ersetzt. Da die Eingabeschicht ein dreikanaliges Bild erwartet, wurde eine CNN-Schicht vor dem Encoder eingefügt, um eine dreikanalige Darstellung zu lernen.

Das Eingabeformat ist $360 \times 640 \times 2$ für Time Surfaces und $360 \times 640 \times 6$ für Event Volumes.

Wir untersuchten weiterhin die Leistung von YOLOv8 [Jocher et al., 2023] auf unserem Datensatz, da es eine State-of-the-Art-Architektur in der framebasierten Bildklassifikation ist. Um YOLOv8 zu verwenden, wurde eine dreikanalige visuelle Darstellung der Zeitflächencodierung mit der von Metavision bereitgestellten Methode erzeugt. Dabei wird die Zeitinformation, die zuvor durch die Grautöne der einzelnen Kanäle dargestellt wurde, in Farbwerte umgewandelt. Diese Bilder wurden dann als Eingabe für das Modell *yolov8m-cls* sowohl zum Training als auch zur Validierung verwendet. Die Gewichte wurden mit Gewichten initialisiert, die zuvor mit dem ImageNet-Datensatz trainiert wurden. Die Visualisierungen werden mit einer Auflösung von 736×736 eingegeben. Die Änderung des Seitenverhältnisses wurde dadurch erreicht, dass die vertikale Achse auf beiden Seiten mit der Hintergrundfarbe der Visualisierung aufgefüllt wird, was bedeutet, dass alle Daten im Originalbild erhalten bleiben, ohne dass etwas abgeschnitten wird. Dies ist wichtig, weil die Klasse von Bildern vollständig von Informationen an den Rändern abhängen kann.

Die von den Modellen erzielten Ergebnisse sind in Tabelle 1 dargestellt. Die Metriken sind definiert als

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

Tabelle 1: Ergebnisse der drei Modelle auf dem Validierungsdatensatz.

Model	Encoding	Accuracy	Precision	Recall
Basic	Time Surface	0.9530	0.9948	0.9158
Basic	Event Volume	0.9875	**0.9967**	0.9795
MobileNetV2	Time Surface	0.9915	0.9876	**0.9965**
MobileNetV2	Event Volume	**0.9955**	0.9949	**0.9965**
Yolov8	Visualization	0.9913	0.9933	0.9901

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

$$Recall = \frac{TP}{TP + FN}. \tag{3}$$

TP(true positive) bezieht sich auf die Anzahl der korrekt klassifizierten Frames, die Personen enthalten, *TN(true negative)* bezieht sich auf die Anzahl der korrekt klassifizierten Frames, die keine Personen enthalten, *FP(false positive)* bezieht sich auf die Anzahl der Frames, die fälschlicherweise als Personen enthaltend klassifiziert wurden und *FN(false negative)* bezieht sich auf die Anzahl der Frames, die fälschlicherweise als keine Personen enthaltend klassifiziert wurden.

Die besten Gesamtergebnisse werden mit MobileNetV2 auf Event Volumes erzielt, dicht gefolgt von MobileNetV2 auf Time Surfaces und Yolov8 auf Time Surface Visualisierungen. Der Baseline-Ansatz bleibt sowohl bei Time Surfaces als auch bei Event-Volumes zurück. Dies zeigt, dass sowohl MobileNetV2 als auch Yolov8 geeignete Kandidaten für den endgültigen Detektor wären. Die Wahl würde hauptsächlich von anderen Faktoren wie der Laufzeitleistung und der verfügbaren Hardware abhängen.

Aufgrund der kürzeren Laufzeit bei vergleichbaren Detektionsergebnissen wurden weitere Tests mit Yolov8 durchgeführt. Um zwischenzeitliche Fehldetektionen auszugleichen, wurden Fenster von 10 kodierten Frames in Folge betrachtet. Alarm wurde nur dann ausgelöst, wenn in 10 aufeinanderfolgenden Frames Personen erkannt werden. Dabei kam es im Zeitraum vom 21.02.2023 bis zum 07.05.2023 zu 2 Fehlalarmen, einmal durch ein in den Tunnel einlaufendes Reh und einmal durch eine in Schrittgeschwindigkeit aus dem Tunnel fahrende Lok. Beide Fälle sind in den Trainingsdaten nicht vorhanden, da sie in den erhobenen Daten nur zu diesen Zeitpunkten auftreten. Mit weiteren passenden Daten wäre eine verlässliche Einordnung dieser Situationen höchstwahrscheinlich möglichst. Gerade die separate Erkennung von eintretenden Tieren wurde von den Projektpartnern als wünschenswerte Funktion identifiziert. Dennoch stellen die 2 Fehlalarme in einem Zeitraum von 3 Monaten eine deutliche Verbesserung gegenüber den laut Schätzung der Bundespolizei ungefähr 10 Fehlalarmen im Monat, welche aktuell ausgelöst werden, dar. Über die tatsächliche Anzahl der Betretungen des Tunnels liegen leider keine genauen Daten vor, alle bekannten Betretungen des Tunnels wurden aber zuverlässig erkannt.

(a) Detektionsergebnis auf Aufnahmen aus dem Testtunnel.

(b) Detektionsergebnis auf selbst erstellten Aufnahmen.

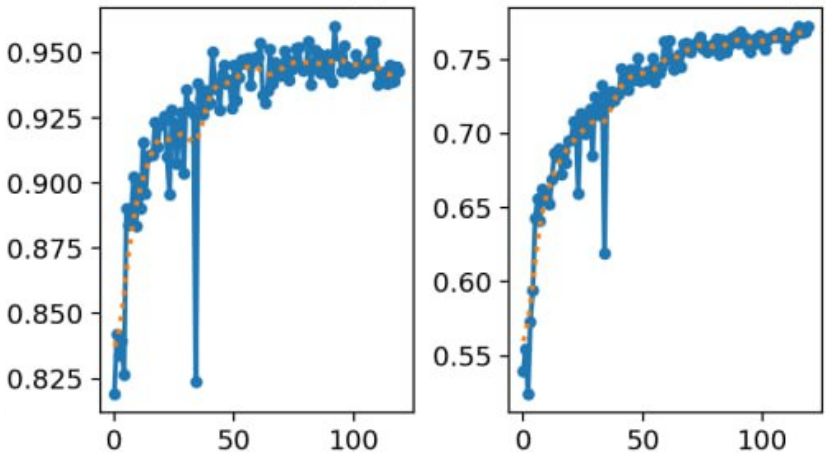
Abbildung 4: Beispiel für Detektionsergebnisse des entwickelten Modells

6.2 Verwendbarkeit von klassichen Frame-Kameras

Das Resultat der Machbarkeitsstudie im Kontext von Frame-Kameras ist ein neuronales Netz basierend auf der YOLOv8-Architektur. Dieses Netzwerk hat sich als ein zuverlässiges Instrument für die Sicherheitsüberwachung in der komplexen Umgebung von Bahntunneln unter den gegebenen Testbedingungen erwiesen. Das erzielte Ergebnis dieses Projektes ist recht zuverlässig, insbesondere unter Berücksichtigung der herausfordernden Lichtverhältnisse, die in diesen Szenarien vorherrschen. Tunnel schaffen eine einzigartige Umgebung, in der das Licht von der Dunkelheit des Tunnels zur Helligkeit des Tageslichts an den Ausgängen stark variiert. Diese Schwankungen führen oft zu Blendeffekten, die die Bilderkennung und -verarbeitung beeinträchtigen können. Einige Beispielhafte Objektdetektionen werden in Abbildung 4 dargestellt.

Das Masasana-Team implementierte mehrere Iterationen des Trainings, um das Modell kontinuierlich zu verbessern und die Genauigkeit der Objekterkennung zu erhöhen. Diese iterative Vorgehensweise ermöglichte es, dass das Modell nicht nur statische Objekte erkennt, sondern auch dynamische Veränderungen im Umfeld, wie die Bewegungen von Menschen und die Geschwindigkeit der Züge, präzise verarbeiten kann.

Die Leistungsfähigkeit des Systems wurde durch Tests bestätigt, bei denen die Erkennungsraten unter verschiedenen Bedingungen bewertet wurden. Diese Tests zeigten, dass die Erkennungsgenauigkeit auch bei sich ändernden Lichtverhältnissen hoch blieb und dass das System effektiv auf die schnellen Wechsel der Lichtverhältnisse reagieren konnte. Die Ergebnisse dieser Tests werden in Abbildung 5 dargestellt. Da es sich jedoch um eine Machbarkeitsstudie handelt und die Datenlage nicht ausreichend hoch war, um alle Use Cases die in der Realität aufkommen können abzudecken, wird im Anschluss der Machbarkeitsstudie empfohlen weitere relevante Objekte mit in die Erkennung aufzunehmen.



(a) mAP50 im Trainingsverlauf (b) mAP50-95 im Trainingsverlauf

Abbildung 5: Darstellung der Detektionsergebnisse im Trainingsverlauf. Die x-Achse stellt jeweils die Trainingsepoche und die y-Achse den Wert der Metrik dar. In der besten Epoche wurden ein mAP50 von $0,945$ und ein mAP50-95 von $0,772$ erreicht.

7 Zahlemäßiger Nachweis

Siehe beigefügtes Formblatt.

8 Notwendigkeit der Geleisteten Arbeit

8.1 Hochschule Niederrhein

Der erste Arbeitsschritt war die Entwicklung der Aufnahmesoftware, welche für die Erfassung und Speicherung der DVS-Daten eingesetzt wurde. Die rudimentären vom Hersteller angebotenen Werkzeuge waren für diesen Zweck aus mehreren Gründen ungeeignet. Die Hauptanforderung an das System war Ausfallsicherheit, damit bei eventuellen Störungen, z.B. in der Stromversorgung oder auch bei unvorhergesehenen Problemen mit der Datenübertragung, die Aufnahme schnellstmöglich fortgesetzt werden konnte, sodass möglichst wenige Daten verloren gehen und am Tunneleingang auftretende zu berücksichtigende Szenarien vollständig erfasst werden. Des weiteren musste die Software mit dem geplanten Austausch der Festplatten umgehen können. Zusätzlich bietet die mitgelieferte Software keine automatisierte Unterteilung der Aufnahme in Abschnitte, wodurch die spätere Weiterverarbeitung erschwert worden wäre. Um diese 3 Anforderungen erfüllen zu können wurde auf Basis der vom Hersteller bereitgestellten Softwarebibliothek eine eigene Aufnahmesoftware entwickelt. Durch diese konnten die Daten über den gesamten Messzeitraum erfolgreich

(a) Der DVS, wie er vom Hersteller geliefert wurde.

(b) Der DVS im angefertigten Gehäuse.

Abbildung 6: Gehäuse des im Tunnel angebrachten DVS

aufgenommen und mehrfach zwischenzeitlich gesichert werden.

Zusätzlich zu der Aufnahmesoftware mussten auch die Anbringung des DVS und die Unterbringung der zur Ausführung der Software benötigten Hardware geplant werden. Da es sich bei dem DVS um ein Evaluationsmodell ohne mitgeliefertes Gehäuse handelte, musste für den Schutz vor Verschmutzung und die Anbringung mit einem geeigneten Aufnahmewinkel ein Gehäuse konzipiert und hergestellt werden. Das Gehäuse musste weiterhin der Sog- und Druckwirkung von durchfahrenden Zügen standhalten. Des Weiteren musste ein Objektiv ausgewählt und beschafft werden, welches auf den Sensor passt und ein geeignetes Sichtfeld gewährleistet. Dank diesem konnte die Aufnahme der Daten unterbrechungslos durchgeführt werden. Das Gehäuse ist in Abbildung 6 dargestellt. Um den Aufbau der Messhardware persönlich miterleben zu können und diese den Anforderungen entsprechend auszulegen, wurde eine Reise zum Testtunnel unternommen.

Parallel zu der Aufzeichnung der Daten am Tunneleingang wurden erste Schritte unternommen, um ein geeignetes Verfahren für die Erkennung von eindringenden Personen zu finden. Zu diesem Zweck wurden zunächst bestehende DVS-Datensätze daraufhin geprüft, ob diese sich für das Training eines Erkennungssystems verwenden ließen. Dies wurde beurteilt, indem auf diesen Datensätzen trainierte Modelle auf selber durch den im Tunnel verwendeten DVS aufgenommenen Daten angewendet wurden. Damit wurde überprüft, ob diese ihre Funktion noch erfüllen konnten. Leider ergab sich, dass sich die bestehenden Datensätze durch Unterschiede in der Auflösung und Aufnahmesituation nicht für unsere Aufgabenstellung anwenden ließen.

Aus diesem Grund wurde entschieden, einen neuen Datensatz zu erzeugen. Zu diesem Zweck wurden weitere gestellte Daten, welche der Situation am Tunneleingang entsprachen, aufgenommen. Dies war notwendig, da zu erwarten war, dass in den am Tunneleingang aufgenommenen Daten nur wenige Personen zu sehen sein würden. Damit das KI-Modell später den Alarmfall erkennen würde, war

eine ausreichende Anzahl an Positivbeispielen notwendig. Es mussten geeignete Szenarien ermittelt, die Aufnahmen erstellt und gelabelt werden. Zeitgleich wurden die ersten vom Aufbau am Tunneleingang erhaltenen Aufnahmen gesichtet und passende Abschnitte als Negativbeispiele in den Datensatz aufgenommen.

Nach der Erstellung eines ersten Datensatzes, der im weiteren Verlauf der Studie immer wieder modifiziert und um vom Tunneleingang erhaltene Daten und weitere gestellte Aufnahmen erweitert wurde, mussten passende Architekturen für das KI-Modell und das Format, in welchem die Event-Daten dem KI-Modell präsentiert werden ausgesucht werden. Für diesen Zweck wurden die in Abschnitt 6 angegebenen Modelle untersucht.

Um den Datensatz weiter zu verbessern und an eine Modellierung der tatsächlichen Umstände heranzuführen, waren Aufnahmen von Menschen neben Zügen und Menschen im Regen notwendig. Die Anforderung, dass Menschen, welche in den Tunnel eindringen während zeitgleich ein Zug ein- oder ausfährt wurde von Seiten der Bundespolizei mehrfach betont. Die Erstellung einer realen Aufnahme von Menschen, welche sich neben einem vorbeifahrenden Zug auf dem Bordstein befinden, wäre für die Darsteller nicht sicher. Aus diesem Grund musste ein Verfahren entwickelt werden, um solche Aufnahmen künstlich aus den bestehenden Daten zu erzeugen. Zusätzlich wurde das Verfahren eingesetzt, um Aufnahmen von Menschen im Regen zu erzeugen.

Nachdem sich Yolov8-cls als das am besten geeignete der ausprobierten KI-Modelle erwiesen hatte, wurde auf Basis dessen ein modellhafter Detektor entwickelt, welcher aus den Vorhersagen des Modells Alarme ausgibt. Dieser Detektor wurde dann auf die gesamten zu diesem Zeitpunkt vorliegenden am Tunneleingang gesammelten Daten angewendet, um dessen Genauigkeit im realen Kontext anstelle von vorher ausgewählten Szenen zu bewerten. Bei diesem Test wurden noch einige Fehlalarme ausgelöst. Dies lag vor allem an Licht das auf nassem Boden reflektierte, aber auch Kleintiere wie Vögel und Insekten lösten Alarm aus.

Auf Basis dieses ersten Tests wurde der Datensatz um Beispiele dieser Situationen erweitert, um die Fehlalarmrate weiter zu reduzieren, wobei die neu trainierten Detektoren jeweils auf den zuvor falsch klassifizierten Situationen und den Validierungsdaten, welche ebenfalls um repräsentative Beispiele erweitert wurden, erneut angewendet wurden. Aus diesen Verbesserungsschritten ergab sich der finale Detektor.

8.2 Masasana

Die Notwendigkeit des durchgeführten Projekts ergibt sich aus der Lücke, die in der Anwendung von neuronalen Netzwerken für spezifische Sicherheitsanforderungen besteht. Zum Zeitpunkt der Initiierung unseres Vorhabens existierte kein ausgereiftes KI-System, das den speziellen Herausforderungen – insbesondere den variablen Lichtverhältnissen und den hohen Sicherheitsanforderungen im Bahnverkehr – gewachsen war. Diese Lücke zu schließen, war von grundlegender Bedeutung, um die Sicherheit in kritischen Infrastrukturen wie Bahntunneln zu erhöhen und die Effizienz der Überwachungsprozesse zu verbessern. Für die

Realisierung des Projekts wurden die zur Verfügung stehenden Mittel gezielt eingesetzt. Ein wesentlicher Teil der Investition floss in das Personal, das für die Entwicklung und Implementierung des neuronalen Netzwerks zuständig war.

Die Steuerung des Vorhabens erforderte ein professionelles Projektmanagement, um die Einhaltung von Zeitplänen, Budgets und die Koordination des interdisziplinären Teams zu gewährleisten. Projektmanagement war unerlässlich, um auf unvorhergesehene Herausforderungen reagieren zu können und die Schnittstelle zwischen technischer Entwicklung, Forschung und den Anforderungen der Deutschen Bahn und der Bundespolizei als Projektpartner zu bilden.

Anstatt eigene Kameras zu installieren, entschied sich die Masasana dafür, auf die bereits von der Deutschen Bahn installierten Kameras zurückzugreifen. Dies war nicht nur kosteneffizienter, sondern ermöglichte auch die Nutzung von hochwertigen, bereits integrierten Systemen, die speziell für den Betrieb in Bahnumgebungen konzipiert waren. Zudem gibt die Nutzung der bestehenden Infrastruktur eine bessere Einschätzung darüber, ob das System auch nach der Machbarkeitsstudie flächendeckend eingesetzt werden könnte, ohne teure und aufwendige Investitionen in Infrastruktur vorzunehmen.

Eine der Kernkomponenten des Projekts war die Generierung eines qualitativ hochwertigen und umfangreichen Datensatzes. Dieser Arbeitsbereich umfasste die Sammlung und Aufnahme von realen Szenarien in den Tunneln, um sicherzustellen, dass das neuronale Netzwerk unter realen Bedingungen effektiv trainiert werden konnte. Die Testdaten mussten eine breite Palette von Szenarien abdecken, um die Robustheit des Systems zu gewährleisten. Letztes konnte leider aufgrund der aufwendigen Streckensperrungen, die für die Erstellung eines Trainingsdatensatzes notwendig sind, nicht ganz so umfangreich durchgeführt werden. Die im Rahmen der Machbarkeitsstudie entwickelten Testdaten lassen jedoch darauf schließen, dass eine Umsetzung und Weiterentwicklung denkbar sind.

Die gesammelten Daten mussten sorgfältig klassifiziert werden, um die Grundlage für das maschinelle Lernen zu schaffen. Jedes Bild oder Video wurde manuell überprüft und annotiert, um sicherzustellen, dass das KI-Modell korrekt zwischen verschiedenen Objekten wie Menschen, Zügen und anderen relevanten Entitäten unterscheiden konnte.

Nachdem das neuronale Netzwerk anfänglich trainiert worden war, war ein kontinuierlicher Prozess der Optimierung erforderlich. Dies umfasste die Feinabstimmung des Modells, um Fehlalarme zu minimieren und die Genauigkeit in schwierigen Lichtverhältnissen zu maximieren. Die Optimierung beinhaltete auch die Anpassung der KI an das Verhalten der Züge und Personen, was für die präzise und zuverlässige Erkennung entscheidend war.

Jedes Arbeitspaket trug wesentlich zum Gesamterfolg des Projekts bei und war dafür ausgelegt die technologischen Zielsetzungen zu erfüllen. Diese strukturierte Herangehensweise stellte sicher, dass jede Phase des Projekts effektiv geplant und umgesetzt wurde, was letztlich zur Entwicklung eines zuverlässigen und effizienten Objekterkennungssystems führte, welches in einer Weiterentwicklung in einen höheren Reifegrad gebracht werden kann.

9 Vorrausichtlicher Nutzen

Auf den Ergebnissen der Machbarkeitsstudie kann in Folgevorhaben aufgebaut werden. Bevor eine praktische Anwendung des Systems in Frage kommen kann, müssen mehrere Punkte weiter bearbeitet werden:

- Zunächst muss ein Test des Detektors über einen langen Zeitraum (≥ 1 Jahr) erfolgen.
- Weiterhin ist eine Anforderungsanalyse bezüglich möglicher Versuche, die Erkennung zu täuschen, notwendig.
  - Darauf basierend ist eventuell ein weiteres Training des Detektors erforderlich.
- Außerdem ist eine differenzierte Erkennung von eindringenden Tieren nötig.
  - Hier ist zwischen Groß- und Kleintieren zu unterscheiden, da das Vorhandensein jeweils eine unterschiedliche Aktion auslösen muss.
- Schließlich ist die Energieeffizienz zu verbessern.
  - Dies beinhaltet sowohl eine energieeffizientere Datenaufnahme als auch
  - eine energieeffizientere Personenerkennung.

Des Weiteren werden die im Rahmen der Machbarkeitsstudie erlangten DVS-Daten der Wissenschaft zur anderweitigen Verwertung zur Verfügung gestellt.

10 Bekannt Gewordene Externe Fortschritte

Iaboni et. al. stellten im Februar 2023 einen $70.75min$ 640×480 Bounding-Box annotierten Datensatz von drohnenbasierten Luftaufnahmen in städtischen Umgebungen, die Fußgänger beinhalten [Iaboni et al., 2023] vor. Dies könnte für die Menschenerkennung in DVS-Daten von Interesse sein. Da das DVS an einer Drohne montiert ist, weist dieser Datensatz aber ähnliche Probleme mit der Eigenbewegung auf wie die vorher beschriebenen Datensätze aus Automobildaten.

Im Bereich der generellen Eindringenserkennung an Bahntunneln gab es nach Projektstart zwei Publikationen, die uns bekannt sind. Diese sind allerdings erst nach Projektende erschienen. Wang et al. untersuchen die Erkennung von Anomalien in Überwachungsmaterial durch die Messung von Rekonstruktionsfehlern in Kamerabildern [Wang et al., 2023]. Li et. al. versuchen, das Problem der seltenen Positivbeispiele bei der Eindringenserkennung zu lösen, indem Hintergrund und Vordergrund von Kamerabildern getrennt werden und das Vorhersagemodell auf Basis von relevanten Teilen des Vordergrunds berechnet wird [Li et al., 2023].

11 Veröffentlichungen

Auf der VISAPP 2023 wurde der NMuPeTS-Datensatz [Bolten et al., 2023] unter dem Titel „N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation“ vorgestellt, welcher unter anderem in Vorbereitung auf dieses Projekt erstellt wurde.

Eine Einreichung eines Beitrages, welcher die Ergebnisse der Machbarkeitsstudie behandelt, ist für eine zukünftige Tagung geplant.

Literatur

[Abadi et al., 2016] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467 [cs].

[Alonso and Murillo, 2019] Alonso, I. and Murillo, A. C. (2019). EV-SegNet: Semantic Segmentation for Event-Based Cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1624–1633, Long Beach, CA, USA. IEEE.

[Benosman et al., 2014] Benosman, R., Clercq, C., Lagorce, X., Ieng, S.-H., and Bartolozzi, C. (2014). Event-Based Visual Flow. *IEEE Transactions on Neural Networks and Learning Systems*, 25(2):407–417. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[Bisulco et al., 2020] Bisulco, A., Cladera Ojeda, F., Isler, V., and Lee, D. D. (2020). Near-Chip Dynamic Vision Filtering for Low-Bandwidth Pedestrian Detection. In *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pages 234–239. ISSN: 2159-3477.

[Bolten et al., 2023] Bolten, T., Neumann, C., Pohle-Fröhlich, R., and Tönnies, K. (2023). N-MuPeTS: Event Camera Dataset for Multi-Person Tracking and Instance Segmentation:. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 290–300, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

[Bolten et al., 2021] Bolten, T., Pohle-Frohlich, R., and Tonnies, K. D. (2021). DVS-OUTLAB: A Neuromorphic Event-Based Long Time Monitoring Dataset for Real-World Outdoor Scenarios. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1348–1357, Nashville, TN, USA. IEEE.

[Catalano et al., 2017] Catalano, A., Bruno, F. A., Galliano, C., Pisco, M., Persiano, G. V., Cutolo, A., and Cusano, A. (2017). An optical fiber intrusion detection system for railway security. *Sensors and Actuators A: Physical*, 253:91–100.

[de Tournemire et al., 2020] de Tournemire, P., Nitti, D., Perot, E., Migliore, D., and Sironi, A. (2020). A Large Scale Event-based Detection Dataset for Automotive. arXiv:2001.08499 [cs, eess].

[Iaboni et al., 2023] Iaboni, C., Kelly, T., and Abichandani, P. (2023). NU-AIR – A Neuromorphic Urban Aerial Dataset for Detection and Localization of Pedestrians and Vehicles. arXiv:2302.09429 [cs].

[Jiang et al., 2019] Jiang, Z., Xia, P., Huang, K., Stechele, W., Chen, G., Bing, Z., and Knoll, A. (2019). Mixed Frame-/Event-Driven Fast Pedestrian Detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8332–8338, Montreal, QC, Canada. IEEE.

[Jocher et al., 2023] Jocher, G., Chaurasia, A., and Qiu, J. (2023). YOLO by Ultralytics (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics.

[Li et al., 2023] Li, B., Tan, L., Wang, F., and Liu, L. (2023). A railway intrusion detection method based on decomposition and semi-supervised learning for accident protection. *Accident Analysis & Prevention*, 189:107124.

[Miao et al., 2019] Miao, S., Chen, G., Ning, X., Zi, Y., Ren, K., Bing, Z., and Knoll, A. (2019). Neuromorphic Vision Datasets for Pedestrian Detection, Action Recognition, and Fall Detection. *Frontiers in Neurorobotics*, 13.

[Mueggler et al., 2015] Mueggler, E., Forster, C., Baumli, N., Gallego, G., and Scaramuzza, D. (2015). Lifetime estimation of events from Dynamic Vision Sensors. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4874–4881. ISSN: 1050-4729.

[Perez-Cutino et al., 2021] Perez-Cutino, M., Eguiluz, A. G., Dios, J. M.-d., and Ollero, A. (2021). Event-based human intrusion detection in UAS using Deep Learning. In *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 91–100, Athens, Greece. IEEE.

[Perot et al., 2020] Perot, E., de Tournemire, P., Nitti, D., Masci, J., and Sironi, A. (2020). Learning to Detect Objects with a 1 Megapixel Event Camera. In *Advances in Neural Information Processing Systems*, volume 33, pages 16639–16652. Curran Associates, Inc.

[Prophesee, 2023] Prophesee (2023). Metavision SDK by Prophesee (Version 4.1.0) [Computer Software].

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252.

[Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, Salt Lake City, UT. IEEE.

[Setola et al., 2015] Setola, R., Sforza, A., Vittorini, V., and Pragliola, C., editors (2015). *Railway Infrastructure Security*, volume 27 of *Topics in Safety, Risk, Reliability and Quality*. Springer International Publishing, Cham.

[Wan et al., 2021] Wan, J., Xia, M., Huang, Z., Tian, L., Zheng, X., Chang, V., Zhu, Y., and Wang, H. (2021). Event-Based Pedestrian Detection Using Dynamic Vision Sensors. *Electronics*, 10(8):888. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[Wang et al., 2023] Wang, Y., Yu, Z., and Zhu, L. (2023). Intrusion detection for high-speed railways based on unsupervised anomaly detection models. *Applied Intelligence*, 53(7):8453–8466.

[Zhu et al., 2019] Zhu, A. Z., Yuan, L., Chaney, K., and Daniilidis, K. (2019). Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion. pages 989–997.