



Fraunhofer

Heinrich Hertz Institute

Abschlussbericht

zur Förderrichtlinie

„Echtzeittechnologien für die maritime Sicherheit“

Projektname

Assistenzsystem zur luftgestützten Videobasierten Echtzeit-Analyse und Objekterkennung mit Hilfe Neuronaler Netzwerke – Hardwarearchitekturen Videocoding und Machine Learning

Akronym

AVALON – ESG

Förderkennzeichen

03SX481E

Förderzeitraum

01.05.2019 - 31.05.2023 (inklusive Uplift) / 31.08.23 (kostenneutrale Verlängerung)

Berichtszeitraum

01.05.2019 - 31.05.2023 (inklusive Uplift) / 31.08.23 (kostenneutrale Verlängerung)

Projektverantwortlicher:

Herr Prof. Dr.-Ing. Benno Stabernack

Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V.

Fraunhofer Institut für Nachrichtentechnik, Heinrich-Hertz-Institut (HHI)

Einsteinufer 37, 10587 Berlin

Tel. +49 30 31002-661

Fax. +49 30 31002-190

benno.stabernack@hhi.fraunhofer.de

Inhaltsverzeichnis

1. Kurzdarstellung.....	3
1.1 Uplift.....	4
2. Ziele und Projektkontext.....	4
2.1 Uplift.....	7
3. Wissenschaftlich-technische Ergebnisse.....	8
3.1 AP 1.2: Erstellung des Lastenheftes.....	8
3.2 AP 1.4: Integration Gesamtsystem.....	8
3.3 AP 4.1.1: ROI Softwaresimulation.....	9
3.4 AP 4.1.2: ROI Hardware Implementierung und Integration.....	9
3.5 AP 4.2.1: Parallelisierungskonzept 4k Videocodec.....	10
3.6 AP 4.2.2: Implementierung Video-Eingangssplitter & Streamsynchronisierung.....	11
3.7 AP 4.2.3: Implementierung ROI Controller für Parallel-Videocodec.....	11
3.8 AP 4.2.4: Gesamtintegration und Simulation Parallel-Videocodec.....	11
3.9 AP 4.2.5: Synthese und Hardwaretests.....	11
3.10 AP 4.2.6: Verlustleistungsoptimierung Videocodec.....	12
3.11 AP 4.3.1: UDP Netzwerkstack.....	12
3.12 AP 4.3.1: RTP Layer.....	13
3.13 AP 2.7 <i>Machine Learning FPGA</i>	13
3.14 AP 4.4.1 <i>Hardwarenahes Simulationsmodell sowie Konzept für ML IP</i>	13
3.15 AP 4.4.2 <i>VHDL Implementierung ML-Coprozessor</i>	13
3.16 AP 4.4.3 <i>Verlustleistungsoptimierung ML-Coprozessor</i>	14
3.17 UPLIFT AP 1: Erarbeiten der Systemanforderungen.....	14
3.18 UPLIFT AP 4.1: Softwarereferenzmodell.....	14
3.19 UPLIFT AP 4.2: Modellierung und Optimierung der ML Beschleunigerarchitektur auf der Basis des in AP 4.1 erstellten Modells.....	14
3.20 UPLIFT AP 4.3: Implementierung der ML Beschleunigerarchitektur.....	15
3.21 UPLIFT AP 4.4: Aufbau einer geeigneten Toolchain zum Training.....	15
3.22 UPLIFT AP 4.5: Benchmarking Nvidia vs. FPGA HW Implementierung.....	15
4. Vergleich des Vorhabenstands mit der ursprünglichen (bzw. mit Zustimmung des ZG geänderten) Arbeits-, Zeit-, und Kostenplanung.....	16
5. Haben sich beim Ziel/Ergebnis bzw. Lösungsweg/ Vorgehensweise Änderungen ergeben?...16	
6. Sind inzwischen von dritter Seite F + E Ergebnisse bekannt geworden, die für die Durchführung des Vorhabens relevant sind?.....	16
7. Fortschreibung des Verwertungsplans.....	16

7.1 Erfindungen/ Schutzrechte.....	16
7.2 Wirtschaftliche Erfolgsaussichten.....	16
7.3 Wissenschaftlich/technische Erfolgsaussichten nach Projektende.....	17
7.4 Wissenschaftliche/wirtschaftliche Anschlussfähigkeit.....	18

1. Kurzdarstellung

Zur Lageaufklärung aus der Luft werden je nach Anforderung Satelliten, bemannte Fluggeräte und seit einigen Jahren auch Drohnen eingesetzt. Sobald eine hohe Auflösung und Echtzeitverarbeitung von Bildmaterial erforderlich ist sind Satelliten ungeeignet und der Einsatz von bemannten Dreh- und Starrflüglern erweist sich als sehr kostspielig. An Land ist heute auch bei Sicherheitskräften der Einsatz von Drohnen zur Aufklärung der Lage aus der Luft nicht mehr neu. Tatsächlich kommen allerdings bei den Behörden und Organisationen mit Sicherheitsaufgaben (BOS) bisher lediglich Kleindrohnen - meist Quadrocopter - zum Einsatz, die üblicherweise mit stark eingeschränkten Möglichkeiten als "fliegende Kamera" mit einer kurzen Flugdauer (typischerweise ca. 20 - 30 min mit entsprechender Nutzlast) und damit nur über kurze Distanzen eingesetzt werden. Auch sind Kleindrohnen/Quadrocopter wetteranfällig für Wind und Regen. Für maritime Einsatzszenarien, aber auch für großflächige Szenarien an Land mit Bezug zu Flussläufen oder Wasserflächen, wie

- Grenzsicherung der Seegebiete
- Überwachung von Seegebieten bzgl. Warenverkehr, Einhaltung von Fischereischutzgesetzen
- Großschadenslagen nach Orkan / Wirbelsturm / Tsunami
- Großschadenslagen nach Havarien i.V. mit gefährlichen Ladungen / Containern
- Überflutungen durch Hochwasser
- großflächige Sturmschäden
- Seerettung
- Überwachung kritischer maritimer Infrastrukturen

usw. sind diese Kleinsysteme ungeeignet. Stattdessen werden z.Z. Hubschrauber und Flächenflugzeuge eingesetzt, was mit deutlich höheren Kosten verbunden ist.

Im Rahmen dieses Vorhabens soll eine innovative Videobildübertragung mit adäquater Sensorik für großflächige Bilderfassung und automatisierter Bildauswertung entwickelt werden, die beispielhaft für fliegende Plattformen auf einer - bisher nur für militärische Zwecke anwendbare - Mittelstreckendrohne mit einer Flugdauer von ca. 12 h für zivile Sicherheitsanwendungen eingesetzt wird.

Im Besonderen wird die Funktionalität „Region of Interest“ (ROI), die hochauflösende Übertragung von mehreren flexibel wählbaren Bildausschnitten erlaubt, wodurch die (Drohnen-) Operateure am Boden signifikant bei den Suchaufgaben unterstützt werden. Es wird angestrebt, die Lösung auch auf anderen fliegenden Plattformen ohne zusätzliche technische Anpassungen einsetzen zu können.

Als Endanwender und Partner des Vorhabens sind insbesondere die BOS des Maritimen Sicherheitszentrums wie die Bundespolizei See, das Havariekommando oder die Wasserschutzpolizeien der Küstenländer relevant.

1.1 Uplift

Das Uplift-Vorhaben ergänzt das bestehende AVALON-System um vier zusätzliche wichtige Merkmale, die dem Endanwender weitere Nutzungspotentiale eröffnen. Einerseits wird die Detektionsrate signifikant erhöht durch die Kombination mit einer oder mehreren Sekundärquellen (z.B. Internet, Seekarten, Sensorik etc.) zur on-board Qualifizierung von Daten vor dem Downstream. Andererseits werden die konzeptionellen Grundlagen für die Skalierbarkeit des Systems und damit dessen Nutzung auf Plattformen mit signifikant unterschiedlicher Payload-Charakteristik zu ermöglichen. Dadurch wird der Einsatz des Systems auf Kleindrohnen (angestrebt 2kg max.) bis hin zu bemannten Plattformen (z.B. Helikoptern) auch in Kombination (Schwarm-Missionen) ermöglicht.

Zu diesem Zweck musste durch das HHI eine signifikante Leistungssteigerung der KI-spezifischen Rechenleistung der Rechnerplattform bei herabgesetzter Verlustleistung realisiert werden:

Durch die HW-mäßige Implementierung aktuellster Verfahren des maschinellen Lernens auf einem FPGA kann die Gesamtrechenleistung des Systems und damit die Erkennungsgenauigkeit hochgradig gesteigert werden. Hierbei soll als Erweiterung des bestehenden Ansatzes, ein sog. Dual-Head Netzwerk verwendet werden, um gleichzeitig nicht nur die Erkennung und Klassifikation von Objekten, sondern auch die dafür notwendige Vorverarbeitung der Videodaten (d.h. die Einteilung in entsprechend klassifizierte Bounding Boxes), vornehmen zu können.

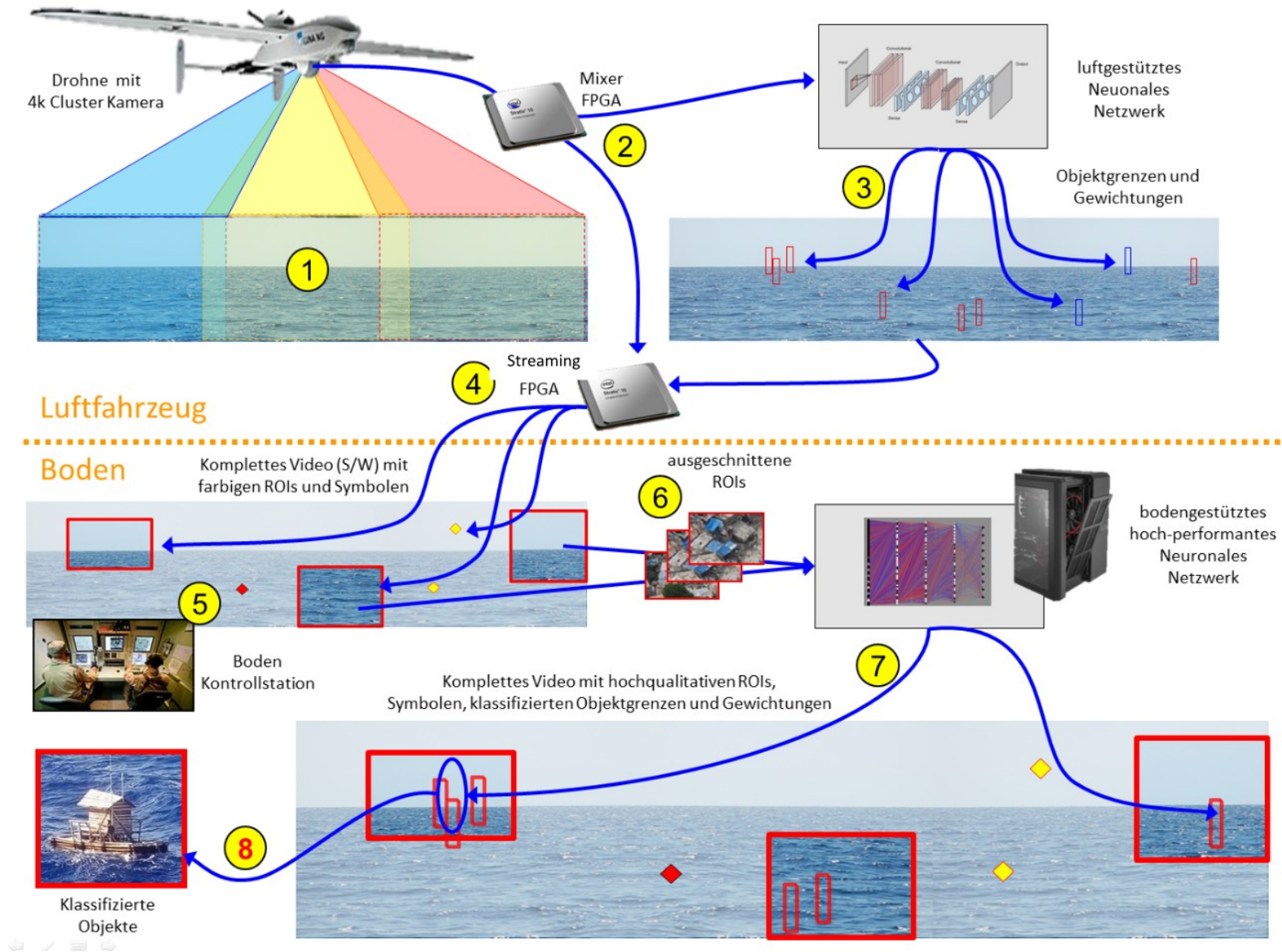
2. Ziele und Projektkontext

Ziel des Projekts ist die Entwicklung einer rekonfigurierbaren Hardwarelösung (FPGA) für die Videoübertragung unter Einsatz innovativer Videokompressionsverfahren mit "Region of interest"-Funktionalität (ROI) und niedriger Latenz. Zur automatisierten Bestimmung der ROI an Bord des Luftfahrzeuges soll der Einsatz eines rekonfigurierbaren Hardwarebausteins (FPGA) für die Implementierung eines neuronalen Netzwerkes mit geringer Verlustleistung evaluiert werden.

Es werden folgende Innovationen **und derzeit nicht am Markt verfügbare Funktionalitäten und Anwendungsmöglichkeiten** für maritime Anwendungen angestrebt:

1. Video-Kompressionsverfahren mit „Region of Interest“ (ROI)-Funktionalität zur Gewährleistung optimaler Bildqualität als notwendige Voraussetzung zur automatisierten Bildanalyse am Boden.
2. Evaluierung des Einsatzes einer rekonfigurierbaren Hardwarelösung (FPGA) zur Auswahl der ROIs mit Hilfe eines neuronalen Netzwerkes.

Abbildung 2 illustriert das Funktions- und Einsatzprinzip des Teilvorhabens (Nummerierung konsistent zum Gesamtvorhaben):



(3) Das Neuronale Netzwerk an Bord der Drohne kann „interessante“ Bildbereiche bzw. Anomalien erkennen und liefert als Ausgabe die Koordinaten von Objektgrenzen („Bounding Boxes“), die gefundene Bildbereiche markieren und die zugehörigen Gewichtungen („Scores“), die die Wahrscheinlichkeit der Vorhersage bewerten. Als Verbesserung einer embedded GPU soll der Einsatz eines verlustleistungsoptimierten neuronalen Netzwerkes auf einem rekonfigurierbaren Hardwarebaustein (FPGA) an Bord des Luftfahrzeuges evaluiert werden.

(4) Die identifizierten Koordinaten und Wahrscheinlichkeiten werden von der Video-Verarbeitungseinheit zur ROI-Definition benutzt. ROIs haben eine größere Fläche als die Objektgrenzen und können mehrere dieser Markierungen enthalten. Anhand der Wahrscheinlichkeitsparameter werden die wichtigsten ROIs direkt in das Video codiert, während weniger wahrscheinliche Bereiche über Symbole mit einem Farbcode markiert werden. Zur weiteren Erhöhung der Kompressionsrate können die Bereiche außerhalb der ROIs optional auch als Graustufenbild übertragen werden.

Die Detailziele des Projektes sind:

(1) Implementierung eines Systems zur signifikanten Verbesserung der Echtzeit-Videoübertragung bei eingeschränkter Kanalbandbreite

Die Qualität der übertragenen Bilddaten wird durch die Nutzung eines weiterentwickelten Videokompressionsverfahren mit "Region of Interest"-Funktionalität und niedriger Latenz verbessert. Die "Region of Interest" (ROI) Technologie dient zur signifikanten Verbesserung der Bildqualität bei akzeptabler Kompressionsrate für die ROI-Bereiche und gewährleistet damit die Echtzeit-Bildanalyse mittels Neuronaler Netze am Boden. Der Inhalt der ROIs wird mit gleichbleibend hoher Bildqualität übertragen und lässt die Auswirkungen schwankender Übertragungsbandbreite nur in den Außenbereichen um das ROI Fenster herum zu. Zur Erfassung weiträumiger Bereiche können mehrere solcher Fokus-Regionen definiert werden und bieten dadurch eine erhebliche Verbesserung bei der visuellen Überwachung. ROIs können vom Operateur der Bodenstation dynamisch in Echtzeit verschoben und in der Größe verändert werden. Im Ergebnis können Sicherheitsorganisationen flexible zu definierende Interessensbereiche in Echtzeit mit Unterstützung von künstlicher Intelligenz auswerten. Die Operateure werden dabei signifikant (bei Entscheidungen) unterstützt.

Messbares Ziel: Datenrate des codierten Videos auf der Funkschnittstelle mit einer maximalen Gesamtdatenrate von 100 Mbit/s sowie einer Latenz des Videosignals von unter einem Frame.

(2) Entwicklung eines automatisierten Analyseverfahrens mit Hilfe neuronaler Netze zum Zweck der luftgestützten Echtzeitbildanalyse und Objekterkennung zur Auswahl der ROI

Zur Bestimmung der ROI im Videobild wird der Videostrom mit Hilfe eines neuronalen Netzes Bilde-weise analysiert. Interessante Objekte, wie z.B. Personen im Wasser, werden erkannt und klassifiziert. Durch die verbesserte Videoqualität im Bereich interessanter Objekte wird der Auswerter am Boden bei der Entscheidungsfindung unterstützt. Das neuronale Netzwerk wird in einem rekonfigurierbaren Hardwarebaustein (FPGA) implementiert. Dadurch können die Anforderungen des Luftfahrzeuges an minimale Verlustleistung, Baugröße und Gewicht erfüllt

werden. Zudem sind Leistungsverbesserungen gegenüber einer GPU-basierten Lösung durch spezifische Anpassungen des Netzes möglich. Die Verbesserungen gegenüber einer embedded GPU werden evaluiert.

Messbares Ziel: geringe Verlustleistung von unter 30 W bei gleichbleibender Erkennungsrate im Vergleich zu GPU-basierter Lösung.

2.1 Uplift

Während der Laufzeit von Uplift soll im Detailziel 3 des Uplift-Vorhabens die Entwicklung einer skalierten Lösung für die Videoverarbeitungseinheit für kleine UAVs durchgeführt werden.

Die derzeit im AVALON - Assistenzsystem für die KI-Funktionalität eingesetzte NVIDIA-Karte weist nicht die notwendige SWAP-C Charakteristika für kleine Plattformen auf.

Kleine Plattformen in diesem Kontext sind insbesondere Drohnen, die von BOS genutzt werden, sich im Größenbereich > 10 kg bewegen und somit professionelle Nutzlast an Bord nehmen können.

Ziel ist die Übertragung der KI-Funktionalität von der proprietären NVIDIA-GPU auf eine Hardware-optimierte und technologieunabhängige FPGA-basierte Lösung. Dies würde den Einsatz des AVALON - Assistenzsystems signifikant auf kostengünstigen COTS-Plattformen ermöglichen.

Zeitgleich soll die Genauigkeit durch Nutzung eines Dual Head DCNN IP Core gesteigert werden.

Beschreibung der technischen Innovation: Aufbauend auf der Entwicklungsarbeit im Projekt AVALON konnte festgestellt werden, dass die notwendige Rechenzeit für die Erkennung und Auswertung entsprechender Videodaten in Hinblick auf die Erkennung Schiffbrüchiger in einigen Punkten wesentlich verbessert werden kann. Die verfolgten Ansätze haben sich zwar als richtig erwiesen, weisen jedoch, auch durch die technologische Weiterentwicklung entsprechender Halbleiter (FPGA), vielversprechendes Entwicklungspotenzial auf. Im Speziellen wird im aktuellen System eine prozessorbasierte Hardware eingesetzt, um eine Vorqualifikation der Daten und damit der Signalisierung der sog. ROI vorzunehmen. Im Rahmen des AVALON Projektes wurden erste Untersuchungen vorgenommen, diese Prozessorplattform durch ein FPGA zu ersetzen, womit einerseits die zur Verfügung stehende Rechenleistung und damit auch die Erkennungsgenauigkeit signifikant gesteigert werden kann, jedoch auch der erforderliche Stromverbrauch gesenkt werden kann. Im Speziellen die Verlustleistungssenkung kommt wiederum dem Ansatz entgegen, skalierbar über unterschiedlichste Plattformen zu sein. Eine vollständige Entwicklung und Systemintegration wurde jedoch im Rahmen von AVALON nicht vorgenommen.

Weiterhin haben sich im Bereich der verwendeten KI Algorithmen ebenfalls Entwicklungen eingestellt, die im Rahmen des Projektes nicht mehr adressiert werden konnten. Der bestehende Ansatz sieht dabei vor, entsprechend vorverarbeitete Videodaten unter Einsatz eines performanten Neuronalen Netzwerkes zu untersuchen und zu qualifizieren. Die hohe Auflösung der Eingangsdaten macht jedoch eine Vorverarbeitung sehr aufwendig in Hinblick auf

die notwendige Rechenleistung. Im Laufe der Projektlaufzeit wurden jedoch sog. Dual Head Netzwerke bekannt, die einerseits die Vorverarbeitung der Eingangsdaten in Hinblick auf Bounding Boxes (ROI) vornimmt, die entsprechend angelernt werden können. Andererseits bieten diese Netzwerke jedoch auch die Möglichkeit die somit gefundenen Objekte innerhalb der Bounding Boxes abermals entsprechend zu qualifizieren.

3. Wissenschaftlich-technische Ergebnisse

Im Folgenden werden die wissenschaftlich-technischen Ergebnisse der Arbeitspakete beschrieben, woran Fraunhofer HHI während der Projektlaufzeit maßgeblich beteiligt war.

3.1 AP 1.2: Erstellung des Lastenheftes

Während der ersten 8 Projektmonate war das HHI in enger Abstimmung mit den Projektpartnern an der Erstellung des Lastenheftes sowie der Systemarchitektur für das Gesamtsystem sowie der Subsysteme und Subsystemkomponenten beteiligt. Dabei könnte das HHI insbesondere seine Expertise für Videocodierung und FPGA-Systeme einbringen.

3.2 AP 1.4: Integration Gesamtsystem

Die Projektpartner wurden umfangreich bei der Integration der gelieferten IP Cores H.264 Videoencoder und Streaming IP (umfasst Transportstrommultiplexer und Netzwerkinterface) in das Gesamtsystem unterstützt. Dafür waren umfangreiche technische Erläuterungen sowie die Erstellung einer Dokumentation erforderlich. Aufgrund der Integration der IP Cores in das Gesamtsystem, wurden einige Verbesserungen insbesondere zum Status des Encoders nötig, welche implementiert wurden.

Die integrierten IP Cores wurden vom Projektpartner Rockwell Collins sowohl im Labormodell als auch im Flugmodell erfolgreich getestet.

3.3 AP 4.1.1: ROI Softwaresimulation

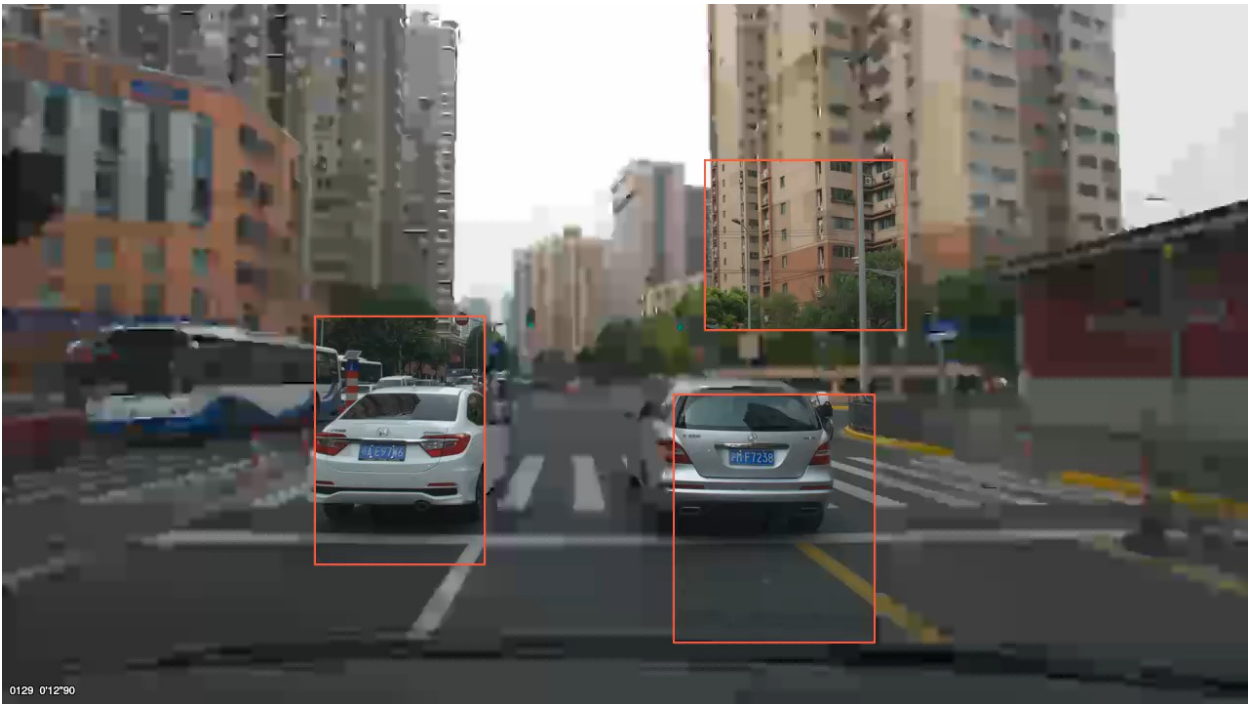


Abbildung 1: Beispiel Videobild mit 3 ROIs

Basierend auf einem existierenden H.264 Softwareencoder wurde die Region-Of-Interest (ROI) Funktion entwickelt. In Abbildung 1: Beispiel Videobild mit 3 ROIs ist ein mit dem Softwareencoder codiertes Videobild mit 3 ROIs zu sehen. Die Bildqualität ist innerhalb der ROI deutlich besser als außerhalb. Mit Hilfe des Softwaremodells wurden Untersuchungen zur Bildqualität in Zusammenhang mit den ROIs durchgeführt. Dabei zeigt sich, wie zu erwarten war, eine starke Abhängigkeit der Bildqualität vom Bildinhalt sowie von der Anzahl und Größe der ROIs.

Der Softwareencoder mit ROI dient als Referenzimplementierung für die Entwicklung eines FPGA-basierten Hardwareencoders mit ROI-Funktionalität.

Der Meilenstein 4.1 wurde mit der Bereitstellung eines Software-Decoders für die Bodenstation fristgerecht erreicht.

3.4 AP 4.1.2: ROI Hardware Implementierung und Integration

Basierend auf den Erkenntnissen des ROI Softwareencoders wurde ein Hardwaredesign auf Basis des vorhandenen Hardwareencoders konzipiert und mit VHDL implementiert. Die Funktionsfähigkeit der Implementierung wurde durch eine funktionale Simulation mit QuestaSim (kommerzielles VHDL Simulationswerkzeug) sowohl für variable als auch für feste Bitraten erfolgreich nachgewiesen. Nach der Portierung auf die Testplattform wurden für die ROI (Region of Interest) Hardwareimplementierung Hardwaretests durchgeführt. Die Funktionsfähigkeit konnte für mehrere ROIs pro Encoder, wie geplant, vollständig festgestellt werden. Der Meilenstein 4.2 wurde somit erreicht.

3.5 AP 4.2.1: Parallelisierungskonzept 4k Videocodec

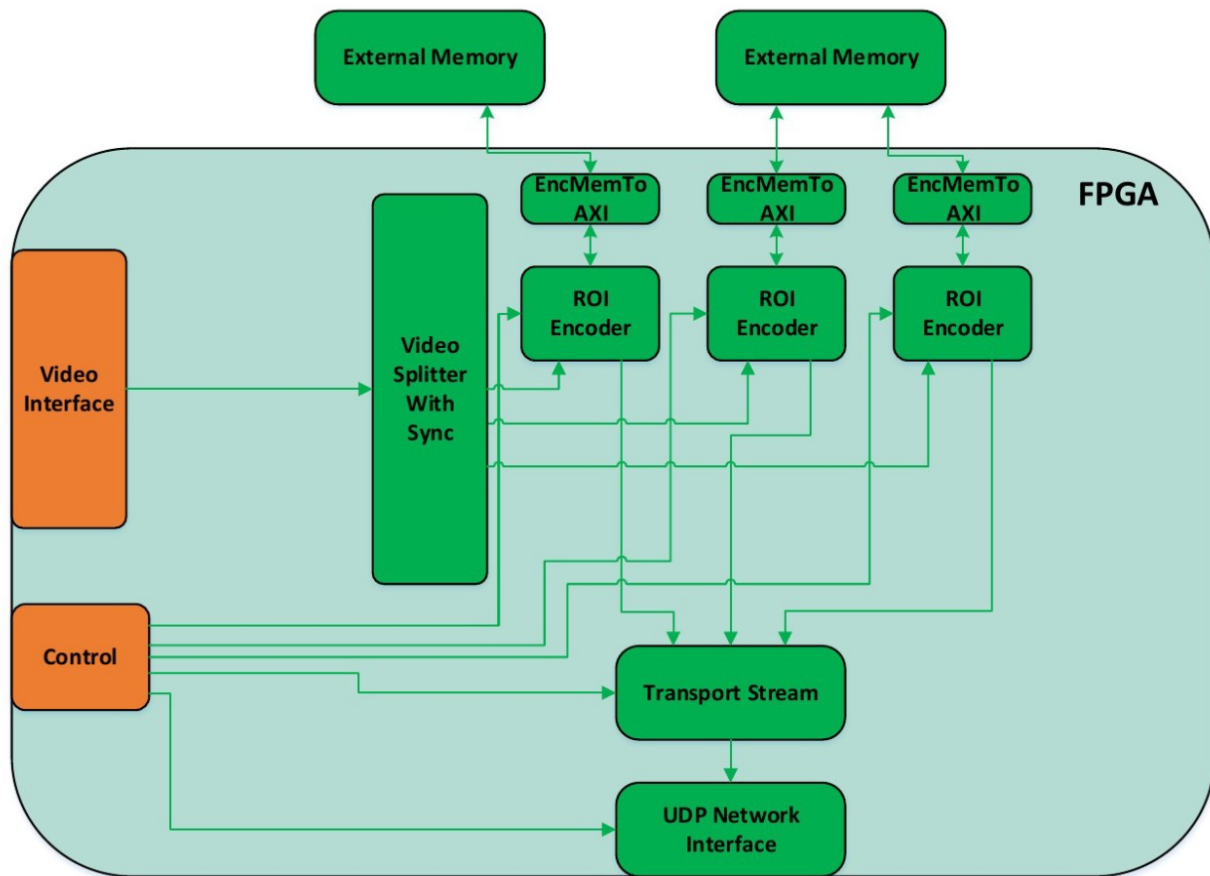


Abbildung 2: Parallelisierungskonzept 4k Encoder

Für die Parallelisierung zur Codierung von 4K Videos wurde ein detailliertes Konzept, welches in Abbildung 2: Parallelisierungskonzept 4k Encoder dargestellt ist, erstellt. Dabei wird das uncodierte 4k Bild streifenweise unabhängig voneinander codiert (s. Abbildung 3: 4K Streifenweise Codierung). Mit Tests von Softwareencoder und -decoder wurde die Funktionsfähigkeit des Konzepts nachgewiesen.

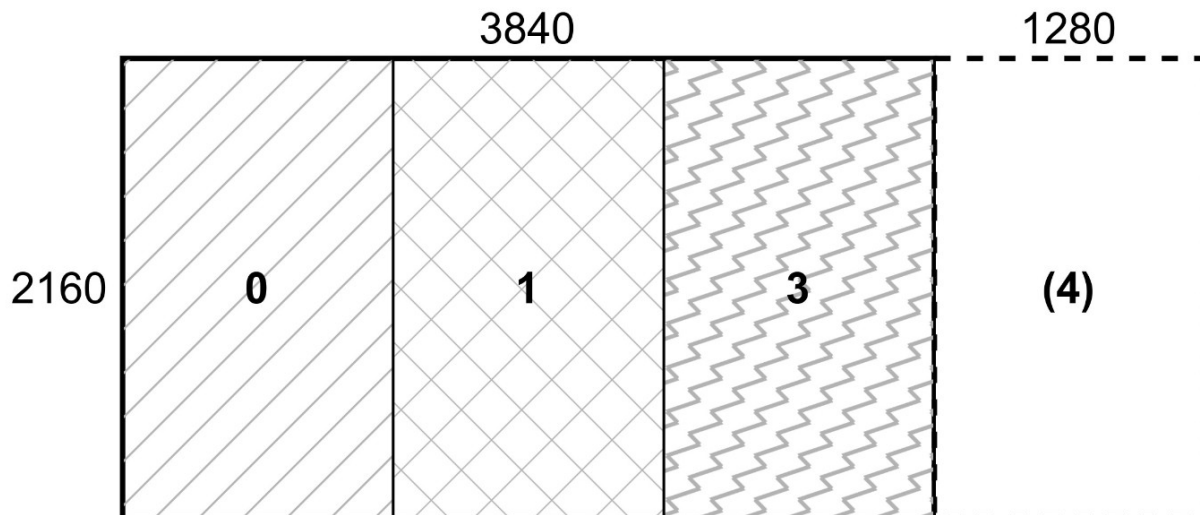


Abbildung 3: 4K Streifenweise Codierung

3.6 AP 4.2.2: Implementierung Video-Eingangssplitter & Streamsynchronisierung

Die im AP 4.2.1 skizzierten Komponenten Video-Eingangssplitter und Streamsynchronisierung wurden in einer Hardwarebeschreibungssprache implementiert sowie in einer separaten Simulation ausführlich getestet. Die korrekte Aufteilung eines 4k Bildes in N Bildspalten ist damit sichergestellt. Auch liegen entsprechende Syntheseergebnisse vor. Die geforderte Taktfrequenz von mindestens 150 MHz wird erreicht. Für Hardwaretest wurden die Komponenten erfolgreich in das Gesamtsystem integriert (s. AP 4.2.4).

3.7 AP 4.2.3: Implementierung ROI Controller für Parallel-Video codec

Durch die Aufteilung des Videosignals auf separate, unabhängige Bildbereiche kann sich eine Region Of Interest (ROI) über mehrere Bildbereiche erstrecken. Im AP wurde eine Kontrolllogik entworfen und getestet, mit welcher die Erzeugung von ROI über mehrere Encoder unterstützt wird.

3.8 AP 4.2.4: Gesamtintegration und Simulation Parallel-Video codec

In diesem Arbeitspaket wurden die Komponenten aus AP 4.2.2 und AP 4.2.3 sowie der existierende Encoder zu einem Gesamtsystem integriert und erfolgreich simuliert. Während eines iterativen Prozess wurden einige Fehler behoben.

3.9 AP 4.2.5: Synthese und Hardwaretests

Im AP 4.2.5 wurde der H.264 Encoder für ein Mittelklasse-FPGA Board für 30 Bilder/sec. synthetisiert. Es folgten erfolgreiche Hardwaretests, zunächst mit wenigen Videosequenzen

und anschließend mit einem umfangreichen Testset von Videos unterschiedlicher Auflösungen, welche mit verschiedenen Konfigurationen (wie Rate Control, ROIs, usw.) codiert wurden.

Die dabei noch gefundenen Fehler wurden iterativ behoben. Zusätzlich wurde der Hardware-Encoder mit dem Streaming-IP zusammengesaltet, um die Funktion der Schnittstellen zu verifizieren. Dies konnte in Hardware erfolgreich nachgewiesen werden. Die Meilensteine 4.4 und 4.5 wurden beide erreicht.

3.10 AP 4.2.6: Verlustleistungsoptimierung Videocodec

Es wurden einige Verlustleistungsoptimierungen durchgeführt. Zusätzlich wurden einige Verbesserungen zur Erhöhung des Taktes implementiert. Umfangreiche Optimierungen konnten aufgrund der nötigen Unterstützungsleistungen für AP1.4 aber nicht durchgeführt werden.

Der Meilenstein 4.6 wurde somit in einem reduzierten Umfang erreicht. Diese wurden aber priorisiert, da sie für die Erreichung des Projektziels wichtiger sind.

3.11 AP 4.3.1: UDP Netzwerkstack

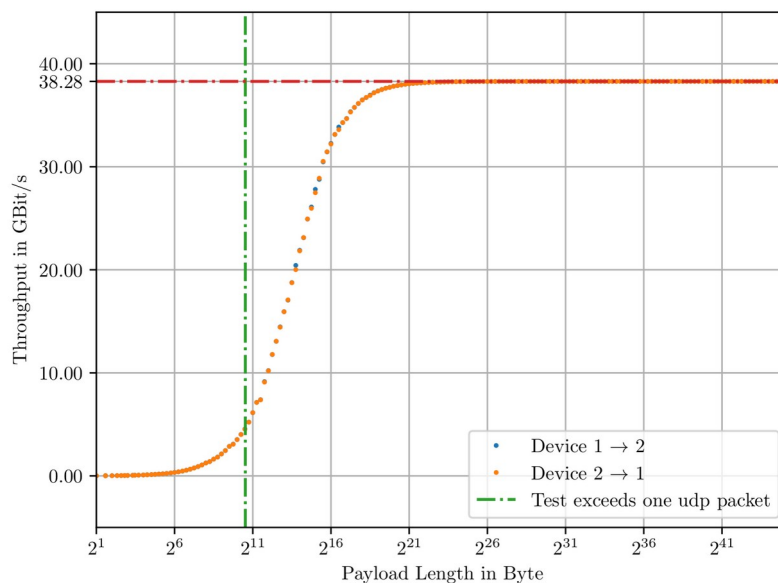


Abbildung 4: Durchsatz UDP/IP Netzwerkschnittstelle

Im Rahmen des Arbeitspaketes wurde eine hochratige, latenzarme UDP/IP Netzwerkschnittstelle basierend auf dem OSI-Schichtenmodell entworfen und in VHDL implementiert. Wie in der Abbildung 4: Durchsatz UDP/IP Netzwerkschnittstelle zu sehen ist, wird gemäß Designziel ein kontinuierliche, hochratiger Datendurchsatz mit bis zu 38,28 Gbit/s Nutzdaten (theoretische Maximalgeschwindigkeit bei 40 Gbit/s nach Abzug der Header) erreicht. Die entwickelte Netzwerkschnittstelle wurde erfolgreich als Hardwaremodul für die Kommunikation zwischen zwei FPGAs sowie zwischen FPGA und PC getestet. Auch die Nutzung eines Switches wurde verifiziert.

Der UDP Netzwerkstack wurde im Berichtszeitraum einigen iterativen Verbesserungen in Bezug auf den nötigen FPGA Ressourcenverbrauch unterzogen. Zudem wurde eine Dokumentation erstellt und der IP Core den Projektpartnern zur Verwendung im Projekt zur Verfügung gestellt.

3.12 AP 4.3.1: RTP Layer

Das geplante RTP (Real-Time Transport Protocol) Protokoll des Streaming IP Cores wurde nach eingehender Recherche in Absprache mit den Projektpartnern durch das MPEG2 (Moving Picture Experts Group) Transportstromprotokoll (TS) ersetzt, da sich diese besser für latenzarme Echtzeitanwendungen eignet. Insbesondere bietet das Transportstromprotokoll ein Mechanismus zur Synchronisation des Taktes des Decoders mit dem Encodertakt.

Im Projekt wurde die TS Komponente in VHDL entworfen, implementiert und erfolgreich getestet. Die TS Pakete werden dann in UDP Pakete gekapselt und über das Netzwerkinterface versendet, was in weiterführenden Test für mehrere HD Videoströme in Hardware getestet wurde. Die Transportströme konnten mit einen Videoplayer (VLC) auf einem entfernten Rechner angezeigt werden.

Der fertige Streaming IP Core mit UDP Netzwerk IP Core und Transportstrom IP Core wurde den Projektpartnern im Meilenstein 4.7 zusammen mit einer Dokumentation zur Verwendung im Projekt zur Verfügung gestellt.

3.13 AP 2.7 Machine Learning FPGA

Im Rahmen dieses AP wurden Überlegungen zur Gestaltung einer Vergleichsstudie zum Einsatz eines Machine Learning FPGAs erstellt. Aufgrund der Praktikabilität wurde die Genauigkeit des ML-FPGA anhand des ILSVRC Testdatensatz, dem FPGA Ressourcenverbrauchs und dem benötigten Energieverbrauch für die Inferenz evaluiert. Trainings- und Testdaten aus dem AVALON Projekt standen für die Entwicklung der FPGA IP Cores jedoch zu spät zur Verfügung.

3.14 AP 4.4.1 Hardwarenahes Simulationsmodell sowie Konzept für ML IP

In dem AP wurde nach intensiver Literaturrecherche ein Konzept für ein FPGA-basierten Machine Learning IP Core zur Erkennung von Personen im Wasser entworfen. Verwendet wird MobileNetV2 aufgrund des günstigen Verhältnisses der Erkennungsgenauigkeit und dem Bedarf an Rechenleistung. Aufbauend auf dem Hardwarekonzept wurde ein hardwarenahes Simulationsmodell erstellt und getestet.

3.15 AP 4.4.2 VHDL Implementierung ML-Coprozessor

In dem AP wurde basierend auf dem in AP 4.4.1 erstellten Simulationsmodell ein ML-Coprozessor in VHDL implementiert. Die einzelnen Netzwerklayer (wie Convolution Layer, Fully Connected, MaxPool) wurden implementiert. Anschließend wurde das Gesamtsystem erfolgreich simuliert und auf dem Ziel-FPGA verifiziert. Es wurde ein echtzeitfähige Implementierung für 30 fps bei 200 MHz Takt erreicht. Durch die Nutzung der 32 bit FP DSPs

konnte eine, gegenüber der Softwareimplementierung, unveränderte Genauigkeit erreicht werden. Der Meilenstein 4.8 wurde somit erreicht.

Aufgrund einer Verzögerung in diesem AP, wurde auf die Optimierung der Verlustleistung im AP 4.4.3 Verlustleistungsoptimierung, wie schon in der Risikoabschätzung des Projektantrags vorgesehen, verzichtet.

3.16 AP 4.4.3 Verlustleistungsoptimierung ML-Coprozessor

Aufgrund einer Verzögerung im AP 4.4.2, wurde auf die Optimierung der Verlustleistung im diesem AP, wie schon in der Risikoabschätzung des Projektantrags vorgesehen, verzichtet.

Der Meilenstein 4.9 wurde deshalb zugunsten des kritischen Meilensteins 4.8 nicht erreicht.

3.17 UPLIFT AP 1: Erarbeiten der Systemanforderungen

Im Rahmen der Aufstockung wurde der Projektpartner RCD durch Zuarbeit in Form von Spezifikationen und bei gemeinsamen monatlichen Statustreffen bei der Erstellung des Lastenhefts umfangreich unterstützt.

3.18 UPLIFT AP 4.1: Softwarereferenzmodell

Es wurde ein Softwarereferenzmodell mit dem verbreiten KI-Framework PyTorch für das Dual-Head Netzwerk YOLOv4-tiny erstellt. Zudem wurde eine VHDL-Testbench unter Nutzung des Testframework vunit erstellt, welche die ML-Parameter direkt aus PyTorch bezieht. Der Meilenstein 4.1 des Uplift-Vorhabens wurde erreicht.

3.19 UPLIFT AP 4.2: Modellierung und Optimierung der ML Beschleunigerarchitektur auf der Basis des in AP 4.1 erstellten Modells

Das in AP4.1 entwickelte Softwaremodell wurde verwendet, um die Architektur hinsichtlich einer möglichen Parallelisierung zu untersuchen. Als Basiskomponente für die Implementierung des ML-Beschleunigers wurde ein im Rahmen des Projektes entwickelter generisch einsetzbarer Rechenkern verwendet, der projektintern als „convengine“ bezeichnet wird. Diese Basiskomponente erlaubt es, mit entsprechender Konfiguration mehrere Layer eines neuronalen Netzes zu berechnen und ggf. auch parallelisiert ausführen zu können. Um das günstigste Verhältnis von implementierten „convengines“ und der verbrauchten Fläche auf einem FPGA zu entsprechendem Datendurchsatz zu ermitteln, wurden unter Einsatz des Softwaremodells entsprechende Untersuchungen durchgeführt. Bei den eingesetzten FPGAs der Firma Intel (ARRIA10) konnte festgestellt werden, dass eine nahezu parallele Implementierung des YOLO-Netzes möglich ist, jedoch der resultierende hohe Datendurchsatz von 600 Inferences/s für die Applikation um den Faktor 10 zu hoch ist. Eine Implementierung unter Einsatz einer sog. convengine erzielt dabei serialisiert schon Durchsätze von 60 Inferences/s .

Wertvoll sind diese Ergebnisse nicht nur hinsichtlich des avisierten Einsatzes mit den festgelegten Parametern, sondern lässt dabei auch Rückschlüsse zu, auf den Einsatz des neuronalen Netzes für Bilddaten mit wesentlich höherer Auflösung, wodurch sich z.B. die Genauigkeit der Erkennung von Menschen in Seenot wesentlich erhöhen ließe.

Speicherbandbreite und Datenformat des Beschleunigers mussten hingegen nicht angepasst werden und konnten durch die Modellierung in ihrer Auslegung bestätigt werden.

Als Ergebnisse dieser Arbeiten konnte planmäßig **MS4.2 als Verifizierter IP Core für Dual Head DCNN in Form eines Softwarereferenzmodells** zur Verfügung gestellt werden.

3.20 UPLIFT AP 4.3: Implementierung der ML Beschleunigerarchitektur

Im Laufe des AP wurde das Softwarereferenzmodell verwendet, um den IP Core für das eigentliche FPGA umsetzen zu können. Hierfür wurde aus der C++ Beschreibung sukzessive ein VHDL Modell erstellt, wobei eine direkte 1:1 Umsetzung nicht möglich war, da die in C++ umgesetzten Strukturen in VHDL eine andere (parallelisierte) Form der Architektur aufweisen.

Neben der Nutzung des Softwaremodells als reine Referenz, wurde das Modell weiterhin eingesetzt, um als Testbench für den im AP umgesetzten VHDL IP Core zu dienen, wodurch die Implementierung komfortabel im Rahmen der angesetzten Zeit umgesetzt werden konnten.

MS 4.3 Trainiertes Modell prototypisch auf HW IP Core getestet konnte somit fristgerecht erfüllt werden.

3.21 UPLIFT AP 4.4: Aufbau einer geeigneten Toolchain zum Training

Im Rahmen dieses Arbeitspakets wurde erfolgreich eine Software-Toolchain basierend auf PyTorch aufgebaut, um beliebige ML-Modelle mittels gelabelter Trainingsdaten trainieren zu können. Für das Training werden aus Effizienzgründen Grafikkarten verwendet.

3.22 UPLIFT AP 4.5: Benchmarking Nvidia vs. FPGA HW Implementierung

Im Rahmen von AP 4.5 wurde vergleichende Tests unter Einsatz der FPGA Implementierung des Beschleunigers auf einem Intel Arria 10 FPGA und einer NVIDIA V100 GPU mit PyTorch Software Implementierung vorgenommen.

Um die Effizienzsteigerungen messbar zu machen, wurden Key-Performance-Indikatoren für den Energieverbrauch definiert. Maßgeblich ist dabei der Energieverbrauch einer Klassifizierung pro Bild bei der Inferenz. Dieser Maßstab stellt die eingesetzte Energie in einem Verhältnis zum Durchsatz (also der Rechenleistung) entsprechend der folgenden Formel:

$$\text{EnergieproFrame} = \frac{\text{Gesamtenergie des Messzeitraums}}{\text{Anzahl der Bilder}}$$

Das FPGA basierte Beschleunigersystem bestehend aus Kontrollserver und Switch ist von allen untersuchten Ausführungsplattformen mit 58,2 mJ/Bild die energieeffizienteste Lösung (vgl.

untenstehende Tabelle). Das FPGA benötigt davon 37,5 mJ/Bild, der Server 13,7 mJ/Bild und der Switch 3,8 mJ/Bild.

Das FPGA basierte Beschleunigersystem ist somit 1.31x effizienter als die Inferenz auf GPUs.

Die Klassifikationsrate sind dabei annähernd gleich gewählt worden.

Plattform	Energieverbrauch pro Bild
GPU System (Server mit Tesla V100)	76 mJ
Davon allein GPU	56 mJ
FPGA Gesamtsystem PCIe-gekoppelt (Bittware 385A) mit Arria 10	74,8 mJ
FPGA NAA Gesamtsystem (Bittware 385A) mit Arria 10	58,2 mJ
Davon allein FPGAs Arria 10	37,5 mJ

4. Vergleich des Vorhabenstands mit der ursprünglichen (bzw. mit Zustimmung des ZG geänderten) Arbeits-, Zeit-, und Kostenplanung

Durch die Verzögerung in AP 4.4.2 konnte AP 4.4.3 leider nicht bearbeitet werden. Allerdings wurden die funktionalen Ziele des ML FPGA IP-Cores unverändert erreicht. Die Kostenplanung ist dagegen unverändert.

Aufgrund der erhöhten Unterstützungsleistung im AP 1.4 fiel der zeitliche Umfang der Verlustleistungsoptimierung des Videoencoders im AP 4.2.6 geringer als geplant aus.

5. Haben sich beim Ziel/Ergebnis bzw. Lösungsweg/ Vorgehensweise Änderungen ergeben?

Von Seiten des Fraunhofer HHI haben sich nach keine wesentlichen Änderungen am Lösungsweg oder der Vorgehensweise ergeben. Auch die Projektziele sind unverändert und wurden erreicht.

6. Sind inzwischen von dritter Seite F + E Ergebnisse bekannt geworden, die für die Durchführung des Vorhabens relevant sind?

Es sind während des Berichtszeitraums keine relevanten Ergebnisse von einer dritten Seite bekannt geworden.

7. Fortschreibung des Verwertungsplans

7.1 Erfindungen/ Schutzrechte

Aktuell sind noch keine Einreichungen von Patenten geplant.

7.2 Wirtschaftliche Erfolgsaussichten

Das Arbeitspaket 4 beschäftigt sich im Wesentlichen mit der Entwicklung von Hardwarekomponenten zur hochrätigen Videosignalverarbeitung. Dies sind hierbei die Kompressionstechnologien, wie auch die Verfahren zur Verarbeitung von Videodaten mit Hilfe

von Machine Learning Ansätzen. Ein wesentlicher Aspekt stellt dabei die echtzeitfähige Verarbeitung unter Berücksichtigung niedrigster Latenz dar. Dieser Aspekt spielt zukünftig eine wesentliche Rolle in Hinblick auf die immer weiter voranschreitende hochratige Vernetzung u.a. auf der Basis des 5G Mobilfunkstandards. Die im Rahmen des Teilprojektes vom Antragsteller entwickelten Technologien stellen z.B. Lösungen für Kernprobleme der Usability (taktiles Internet) dar, wie auch für Anwendungen des Autonomen Fahrens.

Die Arbeiten im Rahmen des AP4 der Aufstockung erhöhen die Erfolgsaussichten der ML-Ansätze durch die Nutzung aktueller Verfahren deutlich.

Die im Rahmen des Projektes erzielten Ergebnisse werden einen wesentlichen Beitrag zur Problemlösung dieser Fragestellungen darstellen.

7.3 Wissenschaftlich/technische Erfolgsaussichten nach Projektende

Die entwickelten Technologien sind eine solide Basis für Veröffentlichungen und Verbesserung des Stands der Technik. Erste Publikationen, welche teilweise im Projekt entstanden sind, wurden bereits veröffentlicht:

- Niklas Schelten, Fritjof Steinert, Anton Schulte and Benno Stabernack, 2020, A High-Throughput, Resource-Efficient Implementation of the RoCEv2 Remote DMA Protocol for Network-Attached Hardware Accelerator, 2020 International Conference on Field-Programmable Technology
- Benno Stabernack and Fritjof Steinert, 2021, Architecture of a Low Latency H.264/AVC Video Codec for robust ML based Image Classification, DASIP 2021: Workshop on Design and Architectures for Signal and Image Processing
- Fritjof Steinert and Benno Stabernack. 2022. Architecture of a Low Latency H.264/AVC Video Codec for Robust ML based Image Classification: How Region of Interests can Minimize the Impact of Coding Artifacts. J. Signal Process. Syst. 94, 7 (Jul 2022), 693–708. <https://doi.org/10.1007/s11265-021-01727-2>
- Uplift: Viktor Herrmann; Justin Knapheide; Fritjof Steinert; Benno Stabernack, A YOLO v3-tiny FPGA Architecture using a Reconfigurable Hardware Accelerator for Real-time Region of Interest Detection 2022 25th Euromicro Conference on Digital System Design (DSD), DOI: [10.1109/dsd57027.2022.00021](https://doi.org/10.1109/dsd57027.2022.00021)
- Carsten Schwarz, Pelin Özkiral, Dr. Jan Pospíšil, Ralf Möllers, Fritjof Steinert, Prof. Dr. Benno Stabernack, Prof. Dr. Andreas Zell, Benjamin Kiefer, Martin Meßmer und Leon Varga, 2023. AVALON - Assistant system for airborne Video-based real-time Analysis and Object recognition using Neural networks im Tagungsband der Statustagung Maritime Technologien 2023. Forschungszentrum Jülich GmbH

7.4 Wissenschaftliche/wirtschaftliche Anschlussfähigkeit

Die in diesem Projekt entwickelten Technologien bieten eine vielversprechende Basis für Anschlussprojekte mit öffentlicher Förderung sowie Direktbeauftragung aus der Industrie. Zudem können die Technologien aufgrund der jahrelangen Erfahrung des Fraunhofer HHI im Form von IP Cores an interessierte Industriekunden lizenziert werden. Eine Lizenzierung der Projektergebnisse an Rockwell Collins ist in Verhandlung.