



HOCHSCHULE RUHR WEST
UNIVERSITY OF APPLIED SCIENCES

INSTITUT INFORMATIK

Technical Report 14-01

VIDEOBASIERTE KAMERAÜBERGREIFENDE BILDSEQUENZANALYSE

U. Handmann, S. Hommel, M. Grimm, D. Malysiak



Technical Report 14-01

Videobasierte kameraübergreifende Bildsequenzanalyse

S. Hommel, M. Grimm, D. Malysiak, U. Handmann

1 INHALTSVERZEICHNIS

2	Einführung.....	3
3	Aufgabenstellung und Voraussetzungen.....	3
4	Wissenschaftlicher und technischer Stand zu Projektbeginn	3
5	Projektergebnisse.....	4
5.1	Spezifikation und Evaluation	4
5.1.1	Systemarchitektur	4
5.1.2	Hardwarekomponenten	6
5.1.3	Testszenarien und Weltmodell.....	8
5.1.4	Videorekorder	8
5.1.5	Integrierte Bildverbesserung.....	9
5.2	Personendetektion und kamerabasiertes Verfolgen	11
5.2.1	Kombination Detektor - Vordergrundsegmentierung.....	11
5.2.2	Merkmalstracker	12
5.2.3	GPU basierte Detektion.....	13
5.2.4	Training des Detektors mit kameraspezifischen Merkmalen.....	16
5.2.5	Fusion multipler Detektoren	16
5.2.6	Nichtlineare Metrik für schwellwertbasierte Detektionsauswahl	17
5.2.7	Personendetektion in parallelen Videodatenströmen.....	18
5.2.8	Evaluation nichtlinearer Kernel	19
5.2.9	Kamerabezogene Ergebnisfusion	20
5.3	Wiedererkennung.....	22
5.3.1	Kamerabezogene Wiedererkennung.....	22
5.3.2	Kameraübergreifende Wiedererkennung	23
5.3.3	Evaluation aller videoindizierenden Verfahren.....	27
5.3.4	SWIR/NIR-Kameras.....	30
6	Abschlussdemonstrator.....	32
7	Literaturverzeichnis.....	33

2 EINFÜHRUNG

Der vorliegende Technical Report beinhaltet die wesentlichen Aspekte des Abschlussberichts des Teilvorhabens „Videobasierte Kameraübergreifende Bildsequenzanalyse“ im Forschungsprojekt „Analyse von Personenbewegungen an Flughäfen mittels zeitlich rückwärts- und vorwärtsgerichteter Videodatenströme (Apfel)“, welches als Verbundprojekt durchgeführt wurde.

3 AUFGABENSTELLUNG UND VORAUSSETZUNGEN

Ziel des Verbundprojektes APFeI (Projektlaufzeit: 01.01.2010 - 31.03.2014, gefördert vom Bundesministerium für Bildung und Forschung) war eine zeitlich vorwärts- und rückwärtsgerichtete Lokalisation von Personen innerhalb eines Kameranetzwerkes aus sich nicht überlappenden Kameras in Hyperechtzeit zu ermöglichen. Einsatzbereiche dieses Szenarios sind kritische Infrastrukturen wie Flughäfen und -plätze. Zunächst fokussierte das Projekt APFeI auf die Lokalisation einer einzelnen Zielperson. Weiterführend wurden die entwickelten Verfahren auf die Analyse von Gruppen erweitert, um Personen als Teil einer Gruppe lokalisieren zu können.

Die Hochschule Ruhr West (HRW) war im ersten Förderungszeitraum hauptverantwortlich für die Spezifikation der Softwarearchitektur und die Evaluation der einzusetzenden Hardware (Rechner, Kameras und Netzwerkkomponenten). Ebenso war die HRW hauptverantwortlich für die Szenarienspezifikation inkl. Sensorlokalisierung und Drehbucherstellung. Orientiert an der Abstraktionsebene der extrahierten Informationen, befasste sich die HRW algorithmisch nah an den Sensordaten. Zu den Aufgaben gehörten die Bildverbesserung und die schnelle Detektion von Personen sowie die Extraktion von Basismerkmalen aus dem Erscheinungsbild von Personen. Des Weiteren war die HRW für die Erstellung geometrischer Modelle der Testumgebungen verantwortlich und unterstützte die TU-Ilmenau bei der temporären Installation von Laserscannern zur anonymen Erfassung statistischer Daten. Abweichend von der Planung zur Projektbeantragung zeigte sich, dass eine Eigenimplementierung kleiner, dezentraler Videoserver einem zentralen System von einem Drittanbieter vorzuziehen ist, da die dezentrale Lösung eine einfache Erweiterung des Systems ermöglicht und damit eine Anpassung an die konkreten Anforderungen des Projektes ermöglicht wurde. Die HRW übernahm die Implementierung und Anpassung der Videoserver. In der Verlängerungsphase vom 01.04.2013 bis 31.03.2014 betrachtete die HRW weiterführend die Detektion des Kopf-Schulter-Bereiches und des Kopfes zur Detektion von Einzelpersonen innerhalb einer dichten Gruppe, sowie die kamerabezogene Verfolgung dieser Bereiche. Besonderer Fokus wurde dabei auf die Optimierung bereits zuvor entwickelter Verfahren in Bezug auf die Gruppenproblematik gelegt. Zudem wurde ein weiterer Klassifikator zur Personenlokalisierung basierend auf Merkmalen des Kopf-Schulter-Bereiches erstellt.

4 WISSENSCHAFTLICHER UND TECHNISCHER STAND ZU PROJEKTBEGINN

Arbeiten die sich mit der kameraübergreifenden Wiedererkennung von Personen bei nicht überlappenden Bereichen im Mittel- bis Fernfeld befassen, lagen zu Beginn des Projektes nicht vor. Dennoch gab es themenverwandte Arbeiten, da für viele Anwendungen im Bereich der Bildverarbeitung das Finden und Verfolgen von Personen von großer Bedeutung ist. Robotik-Anwendungen beispielsweise benötigen häufig die Identität der interagierenden Person und müssen

deshalb eine Erkennung derselben durchführen [1][2]. Andererseits ist es häufig notwendig, den Ort eines Interaktionspartners zu kennen, um diesen zunächst überhaupt erst als solchen zu erkennen oder diesem eventuell zu folgen [1] [3]. Neben eingesetzten Kameras kommen in diesem Umfeld häufig Laser-Scanner zum Einsatz, um den Aufenthaltsort der Personen zu bestimmen, z.B. [4]. Andere Anwendungen, wie assistierende Systeme im Kraftfahrzeug, nutzen bildbasierte Erkennungssysteme, um Gefahrensituationen möglichst früh erkennen zu können [5]. Im Bereich der Sicherheitstechnik liegt der Schwerpunkt der Arbeiten darin, Personen innerhalb von Videobildsequenzen, welche von Überwachungskameras aufgenommen wurden zu detektieren und gegebenenfalls zu verfolgen. Frühe Ansätze fokussierten sich auf eine Bewegungsanalyse (Motion Detection), um das Problem auftretender Nichtrigidität bei Personen zu minimieren. Ziel hierbei ist eine Trennung des Vordergrunds vom Hintergrund, wobei der Hintergrund als bekannt vorausgesetzt wird und gegebenenfalls adaptiert wird, z.B. [6]. Dort werden zusammenhängende Bereiche mit Bewegung zu Personenhypothesen zusammengeführt. Für komplexe Szenarien, insbesondere bei hohem Personenaufkommen, bewegter Kamera oder stark variierendem Hintergrund erweist sich die Modellierung des Hintergrunds als zeitintensiv und die Trennung von einzelnen Personen in Personengruppen als schwierig. Auf der anderen Seite des Verfahrensspektrums werden komplexe Ansätze verfolgt, welche auf lernenden Methoden aufbauen. Beispielsweise wurde von Papageorgiou et al. in [7] ein Personendetektor von Basis Haar-Wavelets und einer Support Vektor Maschine entwickelt. Globale Merkmalsets zur Detektion von Personen finden ebenfalls in verschiedensten Verfahren Anwendung, z.B. kantenbasierte Verfahren [5] [8] und formbasierte Ansätze [9] [10]. Ansätze mit lokalen Merkmalsets (z.B. verwenden Viola und Jones lokale Merkmale basierend auf Haar-Wavelets oder lokaler Bewegung [11] [12] [13] [14]) werden häufig zur Detektion von Personen eingesetzt. Neuere Verfahren, welche eine Personendetektion auf Basis von Kantenorientierungen (sog. HOG-Merkmale - histogram of gradient features) durchführen, liefern vielversprechende Resultate [15] [16]. Ein weiterer formbasierter Ansatz wird von Wu und Nevatia in [17] [18] vorgestellt.

5 PROJEKTERGEBNISSE

5.1 SPEZIFIKATION UND EVALUATION

Eine der wesentlichen Aufgaben der ersten Projektphase bestand im Teilprojekt der HRW in der Spezifikation und Evaluierung der Systemarchitektur und der Hardwarekomponenten, insbesondere der Sensorik, sowie in der Erstellung von Testplänen für die Demonstratorstufen.

5.1.1 SYSTEMARCHITEKTUR

Grundsätzlich können drei Systemkonzepte unterschieden werden, welche eine zentrale, eine dezentrale oder eine hybride Architektur aufweisen. Aus den nachfolgenden Gründen wurde die Systemarchitektur als hybride Lösung umgesetzt.

Eine zentrale Bildspeicherung würde eine sehr hohe Last auf einem Server verursachen, welcher dadurch nur begrenzt um zusätzliche Kameras erweitert werden kann. Hingegen verteilt sich die Last bei einer dezentralen Lösung auf viele lokale Bildserver. Um weitere Kameras einzubinden, müssen lediglich neue lokale Bildserver hinzugefügt werden. Nachteilig bei der dezentralen Bildspeicherung ist, dass nicht jeder Bildverarbeitungsrechner jedes Bild von jeder Kamera über dieselbe Netzwerkverbindung abfragen kann. Dies kann nur durch eine zusätzliche Vernetzung oder einen vorgeschalteten virtuellen Bildserver erreicht werden. Da ein zentraler Bildserver alle Bilder

akquiert, ist eine Synchronisation denkbar einfach. Bei einer dezentralen Lösung hingegen ist hier ein Synchronisationsserver notwendig. Bei der Nutzung eines dezentralen Bildservers erhöht sich allerdings die Ausfallsicherheit, da bei Ausfall einer einzelnen Kamera das Gesamtsystem weiterhin funktionsbereit bleibt.

Eine äquivalente Betrachtung kann für die kamerabezogenen Verarbeitungskomponenten, wie der Bildverbesserung, der Personendetektion und einer kamerabezogenen Personenverfolgung, getroffen werden. Die berechneten Metadaten zum aufgezeichneten Videomaterial können für jede Kamera in dezentrale Datenbanken oder in eine zentrale Datenbank abgelegt werden. Obwohl kleine dezentrale Datenbanken die Skalierbarkeit des Systems vereinfachen, kam in diesem Projekt eine zentrale Datenbank zum Einsatz, da diese einen asynchronen Datenaustausch der Verbundpartner über das Internet während der Entwicklungsphase erleichterte und die Kontroll- und Visualisierungseinheit als zentrale Komponente ausgelegt wurde, welche dem Operator eine Schnittstelle zum System bietet. Weiterhin vereinfacht die zentrale Datenhaltung die kameraübergreifende Analyse im System, wie z.B. Einschränkung des Suchraums auf Basis eines geometrischen Modells der Kameraanordnung. oder die Wiedererkennung von Personen anhand dezentral extrahierter Metadaten der relevanten Kameras, welche zentral in der Datenbank abgelegt sind. Das System wurde am Flughafen Schönefeld in Zusammenarbeit mit dem EASC e.V. als feste Installation realisiert. Weiterhin wurde die Architektur zu Testzwecken temporär auch am Flughafen Erfurt-Weimar umgesetzt [19] [20]. Abbildung 1 zeigt eine Übersicht der Architektur.

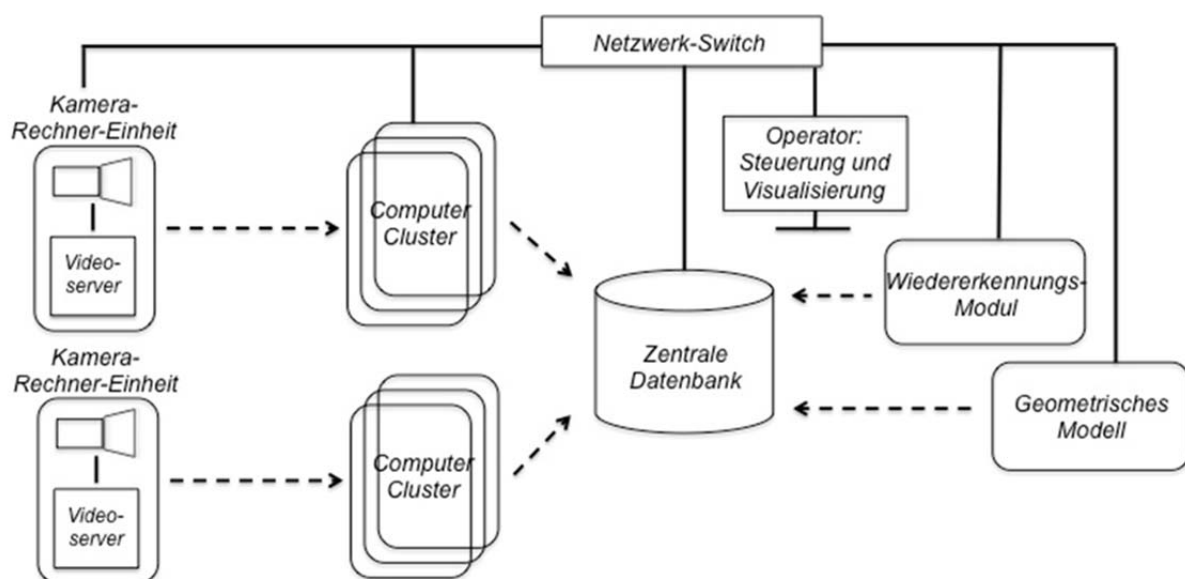


Abbildung 1: Übersicht Systemarchitektur. Dargestellt sind dezentrale (Kamera-Rechner-Einheit, Computer Cluster) und zentrale (Datenbank, Visualisierung, Wiedererkennung, Geometrisches Modell) Komponenten.

5.1.2 HARDWAREKOMponentEN

Um die Rechenlast der verwendeten Algorithmen zu bewältigen, wurden folgende Mindestanforderungen definiert, welche handelsübliche Rechner zum Projektstart anboten:

- Core i7 1,6GHz Prozessor, um je nach Bedarf Algorithmen parallel oder durch Boosting eines Kernes sequenziell in hoher Geschwindigkeit operieren zu lassen
- 8GB-DDR3 Arbeitsspeicher, um große Bildmengen zur optimalen Auslastung der Rechenkapazitäten zwischenspeichern zu können
- 2x1 Gbit/s Ethernetanschlüsse, einen zum Transport der aufgezeichneten Bilddaten und einen zur Kommunikation mit der Datenbank
- die Videoserver müssen zusätzlich über eine mindestens 500GB große Festplatte mit 7200U/min verfügen
- bei dem Videoserver muss eine der beiden Ethernetanschlüsse JumboPackets (MTU=9014Byte) unterstützen, dieser Anschluss wird für die direkte Anbindung der Kamera benötigt

Die Systemspezifikation wurde in der einjährigen Verlängerungsphase des Projektes um performante Grafikkarten, die zur Parallelisierung der Detektion von Personen in stark strukturierten und lebhaften Umgebungen mit großen Personengruppen eingesetzt wurden [19] (siehe Kapitel GPU basierte Detektion), ergänzt.

Anforderungen an die eingesetzten Kameras sind eine hohe Lichtempfindlichkeit, um auch bei schlecht beleuchteten Szenen und in der Dämmerung arbeiten zu können, ein geringes Rauschen, um Störungen bei der Vorverarbeitung und fehlerhafte Merkmale zu vermeiden, eine hohe Dynamik in der Farbe, um möglichst wenig Merkmale zur Wiedererkennung zu verlieren, und eine hohe tatsächliche Auflösung, um entfernte Personen wieder erkennen zu können.

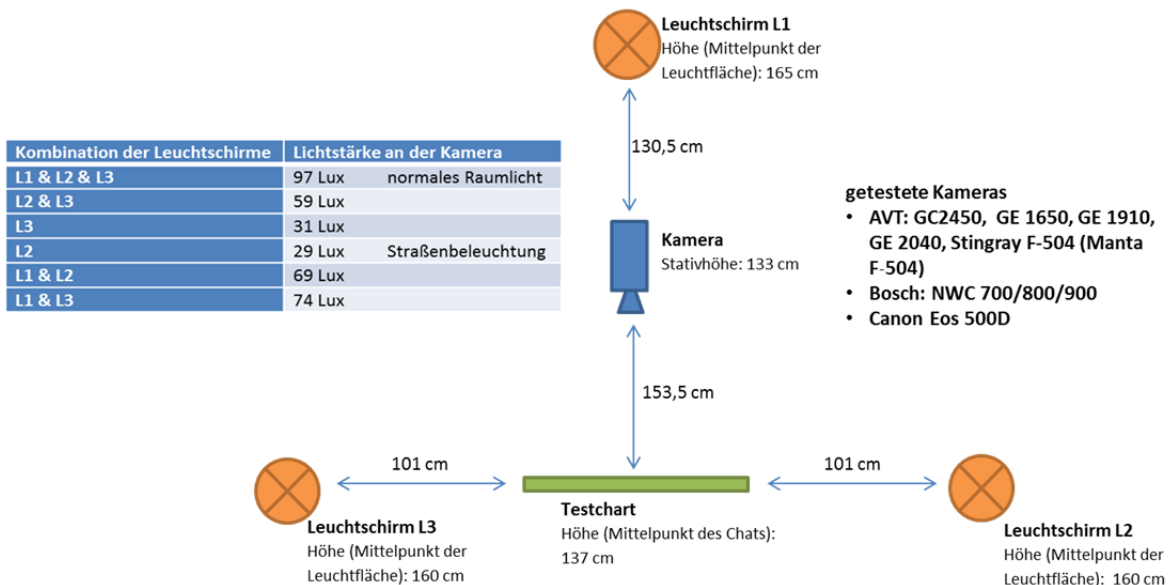


Abbildung 2: Testaufbau für die Kameraevaluation.

Um einen für das Projekt geeigneten Kameratyp zu finden, wurden verschiedene Tests mit dem in Abbildung 2 dargestellten Versuchsaufbau und mit den dort angegebenen Kombinationen der Lichtquellen (L1 - L3) mit mehreren Kameras durchgeführt. Zuerst wurde auf dem standardisierten

Digital ColorChecker SG die Farbdynamik getestet. In Abbildung 3 sind exemplarisch zwei Aufnahmen von unterschiedlichen Kameras zu sehen.



Abbildung 3: Digital ColorChecker SG: Unterschiedliche Farbdynamik verschiedener Kameras.

Auf Basis der durchgeführten Aufnahmen konnte eine ausreichend gute Farbwiedergabe bei folgenden Kameratypen ermittelt werden: Bosch NWC 700, AVT GE 1650, AVT GE 1910, AVT GE 2040 und AVT GC 2450.

Eine Analyse der Weißabgleichfähigkeit wurde bei verschiedenen Beleuchtungen mittels Weißabgleichschart getestet, um Kameratypen zu finden, welche möglichst robust verschiedene Beleuchtungsverhältnisse auf einen vergleichbaren Farbraum übertragen. Weiterhin wurde die Farbdynamik verschiedener Kameratypen betrachtet, da der kamerainterne Weißabgleich Auswirkung auf die Qualität der Farbdynamik hat. In Abbildung 4 ist das Ergebnis verschiedener Kameratypen für differierende Ausleuchtungsszenarien dargestellt. Kameras mit ausreichender Farbdynamik (AVT GE1650 und AVT GE2040) weisen die geringste Standardabweichung und damit den besten Weißabgleich auf.

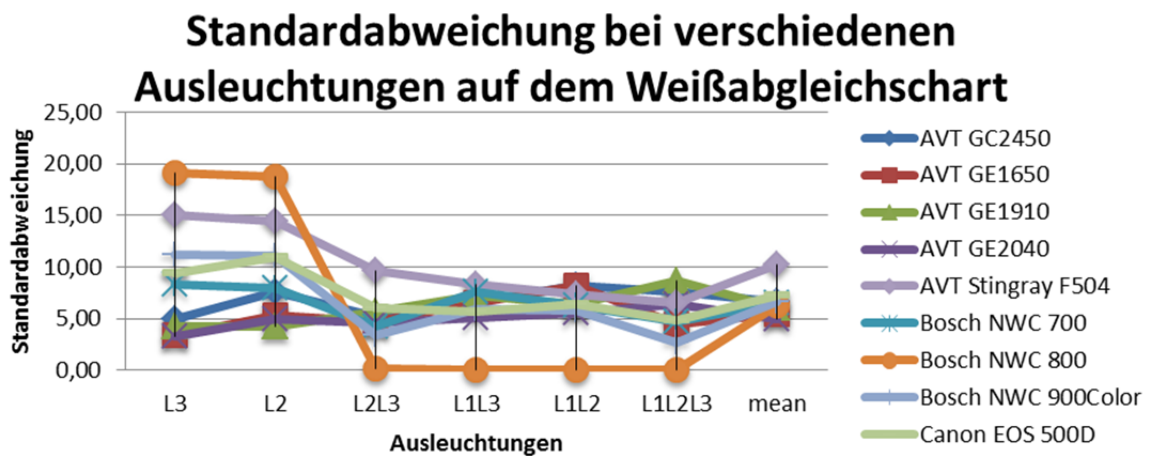


Abbildung 4: Weißabgleich Die Standardabweichung der Farbwerte sollte möglichst gering sein, dennoch müssen alle Farben gut separiert werden.

In einem weiteren Test wurde das mittlere Signal-Rauschverhältnis auf den Farbfeldern des ColorCheckers SG ermittelt. Das Ergebnis ist in Abbildung 5 dargestellt. Die Kameratypen AVT GE1650 und AVT GE2040 weisen hier über alle Beleuchtungssituationen ein sehr gutes Signal-Rauschverhältnis auf.

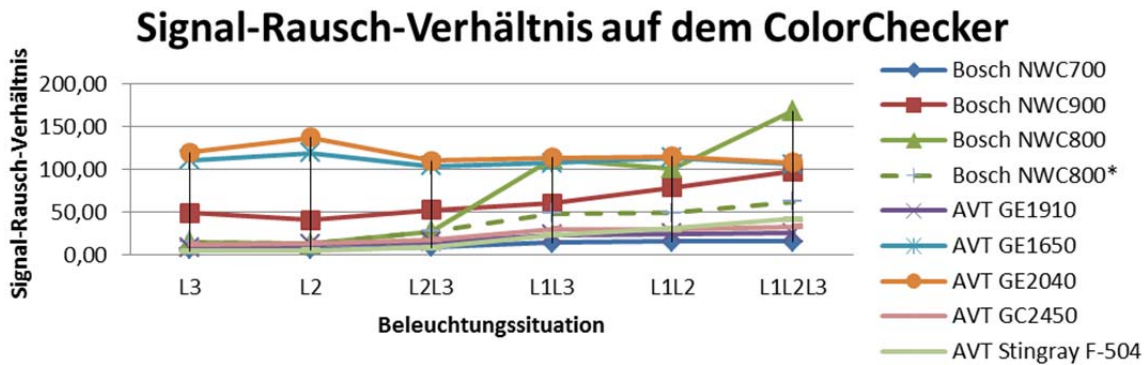


Abbildung 5: Signal-Rausch-Verhältnis: Die Kameras AVT GE1650 und AVT GE2040 zeigen über alle Beleuchtungen ein sehr gutes Signal-Rausch-Verhältnis.

Bei einem letzten Test wurde die Auflösung betrachtet. Hier wurde die Standardabweichung auf feinen Auflösungsrastern bewertet, eine hohe Standardabweichung ist dabei ein Indiz für eine hohe tatsächliche Auflösung. Hierbei schnitten die Bosch NWC 800 und AVT Stingray F-504 am besten ab. Bei dem Kameratyp Stingray liegt dies an einem internen Schärfungsalgorithmus, welcher allerdings zu einem schlechten Signal-Rausch-Verhältnis führt. Ausreichende Qualität erreichen AVT GE1650 und AVT GE2040.

Aus der Gesamtheit aller Test gingen die AVT GE2040 und die AVT GE1650 als die geeignetsten Kameras hervor. Aus Kostengründen wurde sich für die geringer aufgelöste AVT GE1650 entschieden.

5.1.3 TESTSZENARIEN UND WELTMODELL

Die Erstellung von Drehbüchern für Testaufnahmen und die Demonstratoren, sowie die Festlegung der Kamerastandorte erleichtert die Analyse der entwickelten Verfahren. Hierzu wurden am Flughafen Schönhagen und am Flughafen Erfurt-Weimar verschiedene Testsequenzen in verschiedenen Komplexitätsstufen aufgenommen. Beginnend von vereinzelt Personen, welche gezielt in die Kameras schauten, bis hin zu großen Gruppen mit starker Verdeckung bei denen die Zielpersonen gezielt den Blick in die Kameras vermieden. Hierzu wurden typische Laufwege herangezogen (Ergebnisse der Projektpartner TU-Ilmenau und Avistra GmbH).

Im Rahmen der Erstellung der Testszenarien wurden zusätzlich Weltmodelle an den Betrachteten Standorten digitalisiert, welche die Grundlage für kameraübergreifende Analyse bilden (Abbildung 6).



Abbildung 6: Weltmodell: Fotografie EASC (links), erstelltes Modell (rechts).

5.1.4 VIDEOREKORDER

Basierend auf den Hard- und Softwarespezifikationen ergab sich die Notwendigkeit der Implementierung eines eigenentwickelten Videorekorders und -servers. Der Rekorder selbst erfüllt dabei die Funktion, Rohdaten von den Kameras abzugreifen, das Bildmaterial kameranah zu

verarbeiten (Kapitel Integrierte Bildverbesserung) und in zwei Kompressionsstufen abzuspeichern, eine stark komprimierte Version mit VGA-Auflösung zur Visualisierung der Ergebnisse, und eine nur schwach komprimierte Version in HD-Auflösung auf welcher die Verfahren operieren. Außerdem bietet der Rekorder auch Bildserverfunktionalität in dem es über TCP/IP-Verbindungen auf Anfragen die gewünschten Bilder zur Verfügung stellt. Dabei können die stark und die schwach komprimierten Bilder zum aktuellen Zeitpunkt, zu einem bestimmten Zeitpunkt, aus einem Zeitintervall oder dem kompletten verfügbaren Zeitraum angefragt werden. Darüber hinaus kann nach dem Zeitstempel des zuletzt gespeicherten (aktuellen) Bildes und dem Verbindungsstatus gefragt werden. Eine Architekturübersicht ist in Abbildung 7 dargestellt.

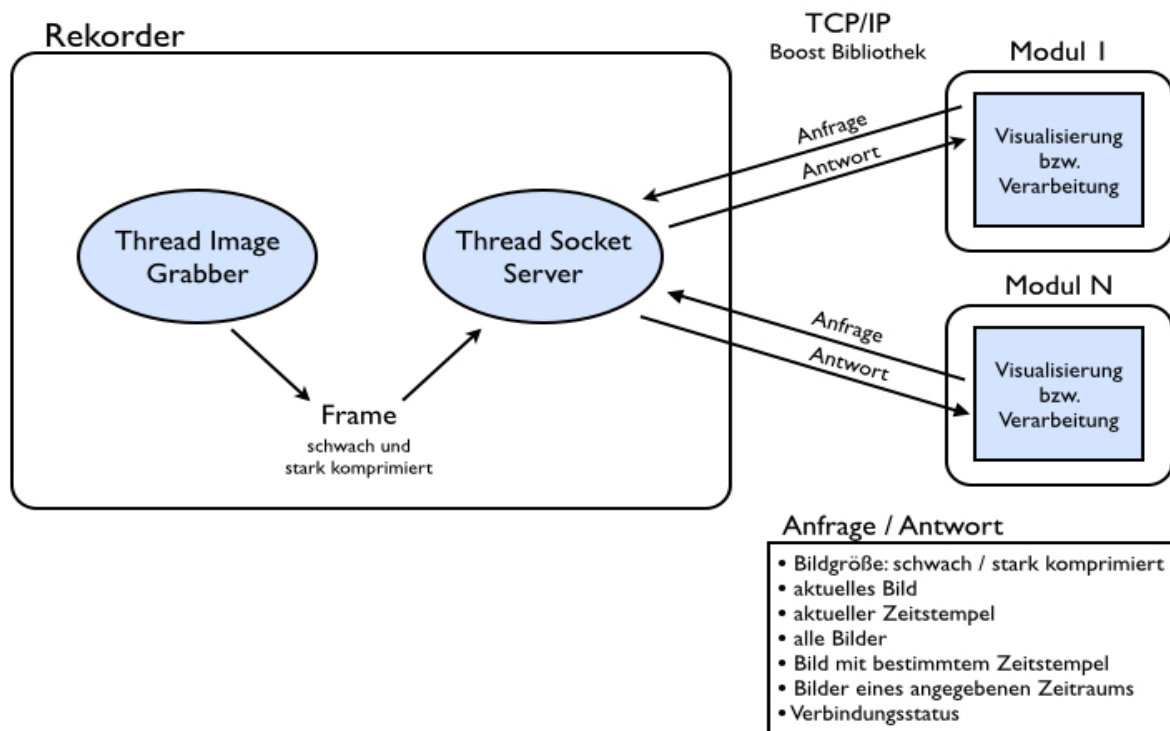


Abbildung 7: Architektur des von der HRW entwickelten Videoservers. Der Videoserver kann mehrere Clients verwalten und auf verschiedene Anfragen reagieren.

5.1.5 INTEGRIERTE BILDVERBESSERUNG

Die integrierte Bildverbesserung dient vor allem dem Ausgleich über und unterbelichteter Bereiche. Um derartige Beleuchtungseinflüsse zu reduzieren wurde eine Bildverbesserung entwickelt, welche die hochaufgelöste kamerainterne Pixelrepräsentation nutzt [19]. Ein Pixel eines gewöhnlichen RGB-Bildes wird durch 3 Kanäle á 8 Bit repräsentiert, wogegen kameraintern ein Farbkanal in der Regel durch deutlich mehr Bit repräsentiert wird, also höheraufgelöste Farbinformationen enthält. Im konkreten Fall arbeitet die eingesetzte Kamera mit 12 Bit je Farbkanal, welche standardmäßig linear auf 8 Bit abgebildet werden. Um überbelichtete Bildbereiche, bei denen der Kamerasensor übersteuert, zu vermeiden, kann die Belichtungszeit reduziert werden. Dies hat den angenehmen Nebeneffekt, dass sich auch die Bewegungsunschärfe reduziert. Dunkle Bildbereiche werden dadurch aber unterrepräsentiert, so dass diese bei einer linearen Abbildung von 12 Bit auf 8 Bit nahezu nicht mehr im resultierenden Bild auszumachen sind. Um diese Bereiche dennoch im 8 Bit-Bild zu erhalten, wird eine günstigere Abbildungsfunktion gewählt, welche diese dunklen Bereiche feiner aufgelöst in das 8 Bit-Bild überführt. In diesem Fall wird eine Gammakorrektur verwendet, wobei $n = 12$ und

Gamma (γ) = 0.63 beträgt. Das Gamma ist so gewählt, dass es minimal ist, aber dennoch die gesamte Breite der 8 Bit Repräsentation ausgenutzt wird.

$$f(x) = (x/2^n)^\gamma * 255, n = \text{Anzahl der kamerainternen bit}$$

Diese „Überrepräsentation“ der dunklen Bereiche bedingt auch eine „Unterrepräsentation“ der hellen, zuvor gar nicht repräsentierten Bereiche. Bei idealem Gamma erhält man so ein Bild mit einer geringeren Bewegungsunschärfe durch Reduktion der Belichtungszeit, und eine gleichmäßige Repräsentation der vormals dunklen als auch überstrahlten Bereiche. Die Annäherung dunkler wie heller Bildbereiche bewirkt so auch den Beleuchtungsausgleich auf seitlich beleuchteten Flächen, beispielsweise dem Gesicht. Da bei einer derartigen Abbildung der 12 Bit auf 8 Bit keine Sensorinformationen dupliziert werden, erhöhen sich auch die Rauschanteile im Bild nicht, wie es beispielsweise bei einem Histogrammausgleich der Fall wäre.

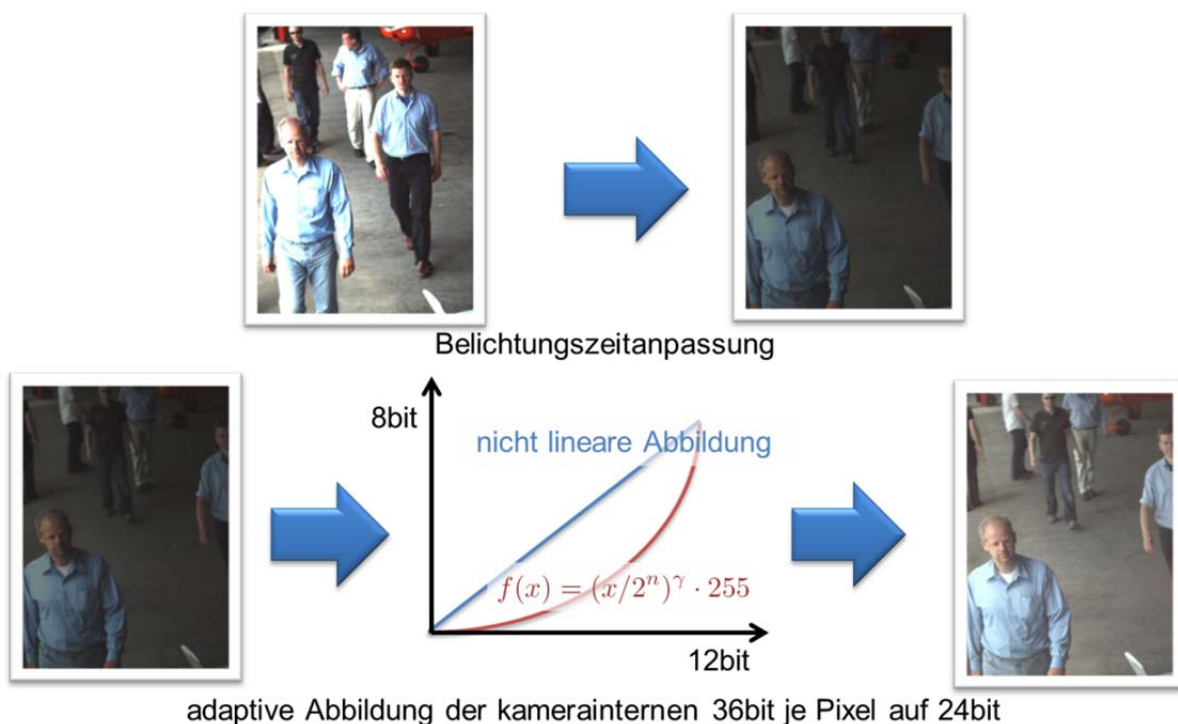


Abbildung 8: Bildverbesserung durch nicht lineare Abbildung.

Um die vorgestellte Bildverbesserung sehr schnell vor der Aufzeichnung der Bilderdaten zu realisieren, wird die Abbildungsfunktion mittels einer vorberechneten Look-Up Tabelle (LUT) realisiert. Die Anpassung an verschiedene Beleuchtungssituationen erfolgt über die Anpassung der Belichtungszeit. Hierzu wird der mittlere Helligkeitswert der Bilder ermittelt und durch langsame Anpassung an einen Richtwert angenähert.

Um dieses Verfahren systematisch testen zu können, wurde ein zweiter Rekorder entwickelt, welcher die 12 Bit Bilder in einem eigenen Format abspeichern kann, um anschließend verschiedene Untersuchungen auf demselben Datenmaterial durchführen zu können. Bei diesen Tests wurde das zuvor empirisch ausgewählte Gamma der Gammakorrektur von 0.63 bestätigt. Hierzu wurden Aufnahmen am Flughafen Erfurt-Weimar, dem Flugplatz Schönhagen sowie den eigenen Räumlichkeiten mit 12 Bit aufgezeichnet. Anschließend wurde das Gamma schrittweise verändert und die Detektionsgüte sowie die Differenzierbarkeit von Farb- und Kanteninformationen getestet

(siehe Kapitel 5.1.). Weiterführend wurde Gamma (γ) anhand der Helligkeitsverteilung im Bild je nach Szenario automatisiert berechnet:

$$\gamma = \frac{\log(\min(\text{Zielwert})) - \log(\max(\text{Zielwert}))}{\log(\min(\text{Eingangswert})) - \log(\max(\text{Eingangswert}))}$$

$$\gamma_{t+1} = \gamma_t + \alpha * (\gamma - \gamma_t); \quad \alpha \in [0,1]$$

5.2 PERSONENDETEKTION UND KAMERABASIERTES VERFOLGEN

Zur Extraktion personenspezifischer Merkmale aus Videobildern ist zunächst eine Lokalisation von Personen in diesen Bildern notwendig. Um die Detektionen zu stützen und zeitlich miteinander zu verknüpfen ist eine kamerabasierte Verfolgung notwendig.

5.2.1 KOMBINATION DETEKTOR - VORDERGRUNDSEGMENTIERUNG

Um die Personenerkennung zu beschleunigen und falsch-positive Detektionen zu reduzieren, ist eine Eingrenzung des Suchraums sinnvoll. In Überwachungsszenarien sind auf Grund der geringen Brennweite der Kameras oft große Bildbereiche unbelebt. Aber auch zeitlich betrachtet gibt es öfters Bereiche ohne Aktivität.

In Videoaufnahmen mit einem nur kleinen interessanten Bereich mit temporärer Veränderung eignen sich Verfahren wie eine Vordergrundsegmentierung oder eine Bewegungsdetektion um den Suchraum nach Personen einzuschränken. Beide Verfahren wurden in Kombination mit einem HOG¹-Personendetektor getestet und führen zu einer massiven Beschleunigung der Detektion. Da sich nicht alle Personen bewegen ist die Bewegungsdetektion nur begrenzt einsetzbar. Die Vordergrundsegmentierung muss allerdings adaptiv gestaltet sein um sich an veränderte Bedingungen anpassen zu können. Um hierbei keine Personen zu verlieren, wird das Hintergrundmodell nur in den Bereichen adaptiert, an denen keine Person detektiert wurde. Als schnelle Vordergrundsegmentierung wird hier die Differenzbildanalyse eingesetzt. Bei dieser muss zur Berücksichtigung von Rauschen eine Schwelle für die pixelweise Differenz zwischen Hintergrundbild und der aktuellen Aufnahme bestimmt werden, um möglichst wenig Rauschen als Vordergrund zu klassifizieren, aber auch um keine Vordergrundbereiche zu verlieren. Hierzu wird die Standardabweichung der Helligkeitswerte des Bildes verwendet [19].

Die Kombination beider Verfahren (siehe Abbildung 9) erhöht die Verarbeitungsgeschwindigkeit des Detektors, da nicht nur der Suchraum auf dem Bild verkleinert wird, sondern sich auch die Suche über verschiedene Skalierungsstufen des Personenmodells entsprechend reduzieren. Auf Grund zeitlicher Lücken ohne Detektionen, können in dieser Zeit sich eventuell noch im Bildspeicher befindlichen Aufnahmen abgearbeitet werden.

¹ Histogram of Oriented Gradients

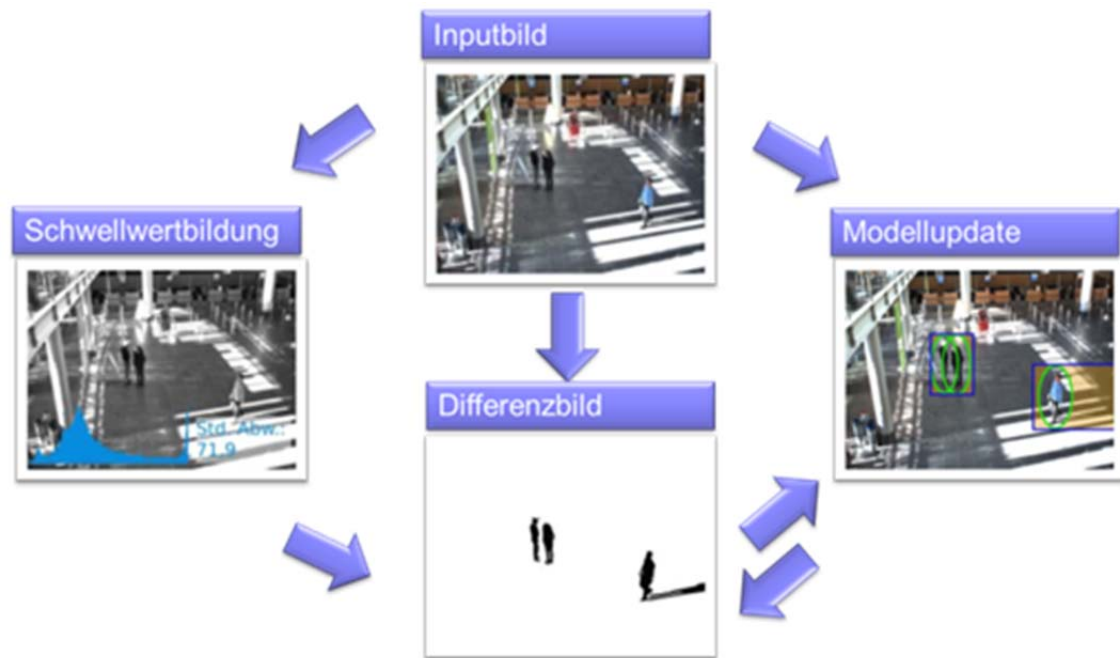


Abbildung 9: Kombination aus Personendetektor und Vordergrundsegmentierung: *Personen werden nur im Vordergrund gesucht und der Hintergrund wird an Personendetektionen nicht aktualisiert.*

5.2.2 MERKMALSTRACKER

Um eine spätere Wiedererkennung von Personen zu beschleunigen und fehlende Personendetektionen zu kompensieren, werden bereits während der Liveanalyse erste Merkmale aller detektierten Personen extrahiert und kamerabezogen verfolgt. Die kamerabezogene Verfolgung von Merkmalen ermöglicht die schnelle Zuordnung von zeitlich aufeinanderfolgend extrahierten Merkmalen zu einer Person. Dies erlaubt die effiziente Erstellung von Merkmalsräumen einer zu suchenden Person. Hierzu werden die zwischen den Personendetektionen extrahierten Merkmale verfolgt und mit neuen Personendetektionen verknüpft. Während dieser Verfolgung von Merkmalen wird die Ähnlichkeit zu einer eingangs ermittelten Referenz bestimmt und zusammen mit den Merkmalen auf der Datenbank abgelegt. Sinkt die Ähnlichkeit unter einen Schwellwert, so wird die Verfolgung dieser Merkmale abgebrochen. Dadurch wird verhindert, dass Merkmale fälschlich vom Hintergrund extrahiert oder Merkmale verschiedener Personen verknüpft werden, sobald sich Personen aus dem Bildbereich bewegen. Auch bei einer starken Änderung der Ansicht einer Person bricht so die Verfolgung der Merkmale ab, sobald diese nicht mehr ausgemacht werden können. Dadurch entsteht, sobald eine neue Personendetektion vorliegt, ein neuer Track. Die zu verfolgenden Merkmale werden aus drei Bereichen innerhalb einer Personendetektion ermittelt (Abbildung 10). Aus einem rechteckigen Bereich des Unter- und Oberkörpers werden der mittlere Farbwert und die mittlere Intensität der Farbe extrahiert [21]. Dies ermöglicht eine erste Unterscheidung von Personen, dessen Bekleidung sich bereits durch deren Grundfarbe unterscheiden. Aus dem Bereich des Oberkörpers werden zudem die Anteile an horizontalen und vertikalen Kanten bestimmt. Hierdurch lassen sich Personen mit stark texturierter Kleidung von Personen mit homogener Kleidung unterscheiden. Darüber hinaus ist auch eine Aussage über die Richtung der Textur möglich, wodurch z.B. Personen mit kariert, längs- und quergestreifter Kleidung untereinander unterschieden werden können.



Abbildung 10: Merkmalstracker: Die zu verfolgenden Merkmale werden aus den hier dargestellten Bereichen extrahiert.

Die Extraktion von Texturmerkmalen aus dem unteren Rechteckbereich ist hingegen nicht zielführend, da dort die Textur primär von der Faltenbildung der Kleidung bestimmt wird. Ein dritter Bereich beschreibt ein Oval auf dem Oberkörper. Aus diesem Bereich wird je ein normiertes 16-wertiges Histogramm über die Farbwerte und die Intensitätswerte erstellt. Dies repräsentiert verschiedene Farbverteilungen der Kleidung bei z.B. gestreifter Kleidung, aber auch besondere Feinheiten wie ein sich farblich abhebende Logos oder eine Krawatte und Ähnliches.

5.2.3 GPU BASIERTE DETEKTION

Zuverlässige und robuste Algorithmen für die Detektion von Personen, wie das HOG-Verfahren, benötigen sehr viel Rechenzeit. Um diese zu reduzieren kommen Kombinationen wie bereits beschrieben zum Einsatz, auch die Kombination mit Tracking-Verfahren um fehlende Detektionen zu kompensieren ist denkbar. Die Einschränkung des Detektionsraumes führt in belebten Situationen mit vielen Personen allerdings nicht zu der erhofften Beschleunigung, da hier über längere Zeiträume nahezu das gesamte Bild betrachtet werden muss. Eine Detektion auf der spezifizierten CPU benötigt auf den HD Aufnahmen (1600px * 1200px) ca. 3,5s pro Bild. Kompensiert man in dieser Zeit die fehlenden Detektionen mit einem Tracking-Verfahren können neu in den Kamerabereich kommende Personen nicht erfasst werden, auch Personen die stark verdeckt werden gehen verloren und können in dem Zeitraum bis zur nächsten Detektion nicht erneut betrachtet werden.

Eine weitere Möglichkeit die Detektion zu beschleunigen ist die Parallelisierung des Detektionsverfahrens. Hierfür, hat sich die Hochschule Ruhr West dazu entschieden, die Rechenleistung von Grafikkarten zu nutzen. Im Gegensatz zur CPU verfügt eine GPU über eine Vielzahl an Rechenkernen. Dies ermöglicht eine massive Parallelisierung und führt somit zu einer stark erhöhten Rechenleistung. Dadurch steht ein universell einsetzbares Verfahren zur Verfügung, welches nicht nur auf spezifischen Regionen oder unter Annahme typischer Personengrößen arbeitet. Für eine erste Programmierung wurde das CUDA-Framework von NVIDIA verwendet. Es kamen zwei NVIDIA GeForce GTX 590 Grafikkarten je Kamera zum Einsatz, wobei jede Grafikkarte über zwei GPUs mit jeweils 512 CUDA Recheneinheiten verfügt. Die Rechenzeit dieser ersten Umsetzung beträgt 350ms pro Bild, wobei alle 100ms ein neues Bild aufgezeichnet wird [22]. Mit den eingesetzten 4 GPUs ist somit eine videotaktschritthaltende Detektion von Personen möglich (Abbildung 11).

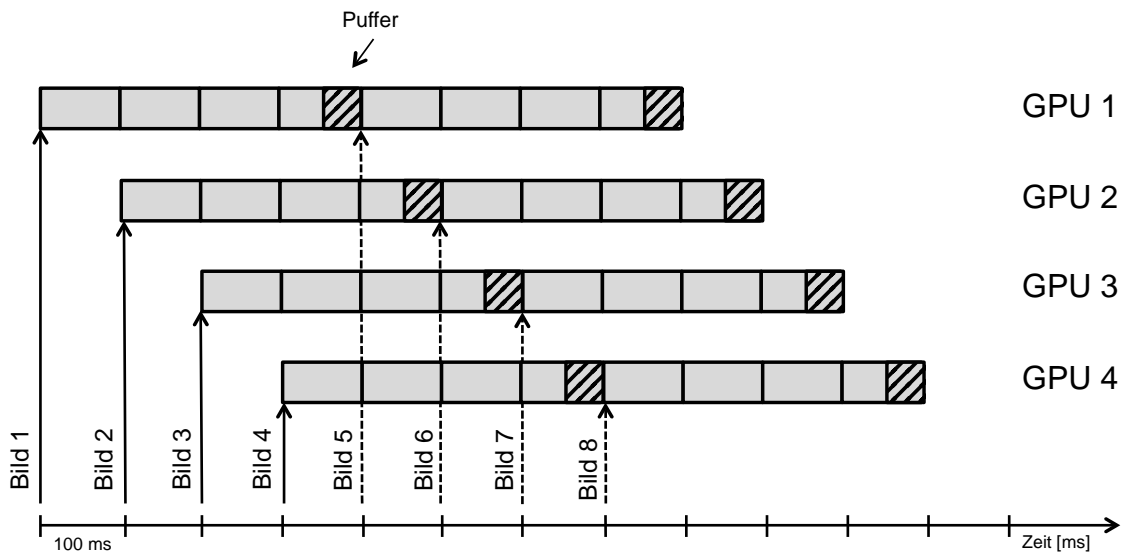


Abbildung 11: Verarbeitungszyklus auf den GPUs. Alle 100ms wird ein neues Bild zur Verarbeitung an eine freie GPU gesendet. Jede GPU benötigt ca. 350ms pro Bild, so dass bei vier GPUs eine videotaktische Verarbeitung erreicht wird

Um die Problematik verdeckter Personen berücksichtigen zu können, kommen zwei weitere Detektoren zum Einsatz, ein Kopf-Schulter- und ein Kopfdetektor. Durch Einbeziehung in einen Fusionsansatz [22] lassen sich damit auch Personen detektieren, bei denen eine Ganzkörperdetektion fehlschlägt. Mit dem oben beschriebenen Ansatz ist allerdings keine parallele Detektion aller drei Bereiche mittels der 4 GPUs möglich. Durch Optimierung des Algorithmus konnte die benötigte Zeit aber auf 140ms je Bild und Detektor reduziert werden. Durch Erweiterung um eine weitere Grafikkarte ist so eine parallele Verarbeitung möglich (Abbildung 12). Es können sowohl alle drei Verfahren für eine Kamera auf einem Rechner operieren, oder die Bilder von drei Kameras mit nur einem GPU-Rechner durch ein Verfahren analysiert werden.

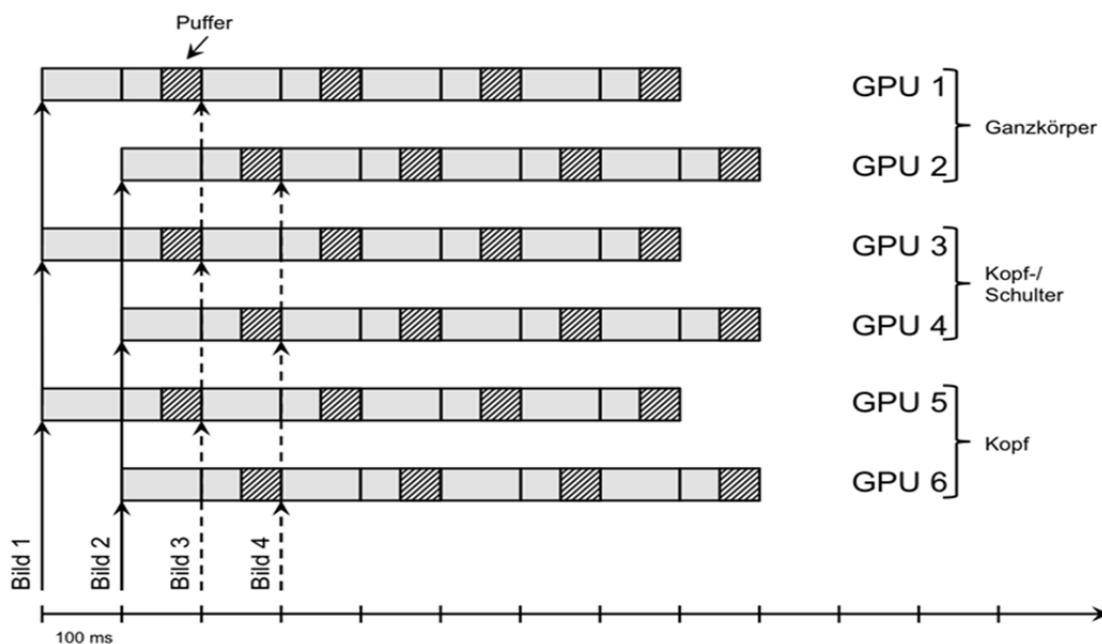


Abbildung 12: Verarbeitungszyklus auf den GPUs. Jeweils zwei GPUs werden verwendet, um ein Detektionsverfahren videotaktisch auszuführen. Dabei wird ein Bild von drei GPUs mit unterschiedlichen Klassifikatoren gleichzeitig verarbeitet.

Da die hier verwendete Konfiguration eines GPU-Rechners erheblich von einer Mainstream-Konfiguration abweicht, wurden im weiteren zusätzliche Optimierungen vorgenommen.

Zunächst wurde der Einsatz neuerer GPUs getestet. Dazu wurde im Hinblick auf eine weitere Effizienzsteigerung der GPU basierten Detektionsverfahren die implementierten Algorithmen auf Grafikkarten der Nvidia Kepler-Reihe getestet. Die Kepler Architektur stellt den direkten Nachfolger zur Fermi Architektur dar, zu welcher die bereits eingesetzte GTX580 gehört. Der Test bestand aus einer Evaluation der Algorithmen auf Grafikkarten der GTX670 und GTX680 Reihe. Hierbei stellte sich heraus, dass trotz verbesserter physikalischer und struktureller Eigenschaften der GPUs, die tatsächliche Laufzeit der Algorithmen nicht verbessert wurde. Eine genauere Untersuchung der Resultate führte zu der Feststellung, dass die Kepler-Reihe für GPU-basierte Berechnungen nicht optimiert wurde bzw. ihre verbesserten Eigenschaften in diesem Anwendungsbereich nicht zum Tragen kommen (Abbildung 13). Aufgrund dieses Umstands wurde nach alternativen Plattformen gesucht, auf welchen die Detektionsalgorithmen effizient ausgeführt werden können. Dazu wurden elementare und gut parallelisierbare Algorithmen auf diversen GPUs hinsichtlich ihrer Ausführungsdauer getestet. Diese elementaren Algorithmen dienen als plattformübergreifendes Maß für die Leistung einer GPU. Hierbei zeigten AMD GPUs der Reihen Cayman und Tahiti (aktuelle Generation) ein wesentlich größeres Leistungspotential als vergleichbare Nvidia Produkte. Dieses Resultat indiziert deutlich den Einsatz von AMD GPUs, jedoch ist eine direkte Ausführung der bereits implementierten Algorithmen nicht ohne weiteres möglich. Die Implementierungen liegen in der Programmiersprache CUDA vor, hierbei handelt es sich um eine Nvidia-spezifische Erweiterung der Sprache C. AMD GPUs werden jedoch in der plattformübergreifenden Sprache OpenCL angesprochen, welche ebenfalls eine Erweiterung der Sprache C darstellt. Zur Nutzung der Leistung einer AMD GPU wurde somit eine Portierung der bestehenden Algorithmen durchgeführt.

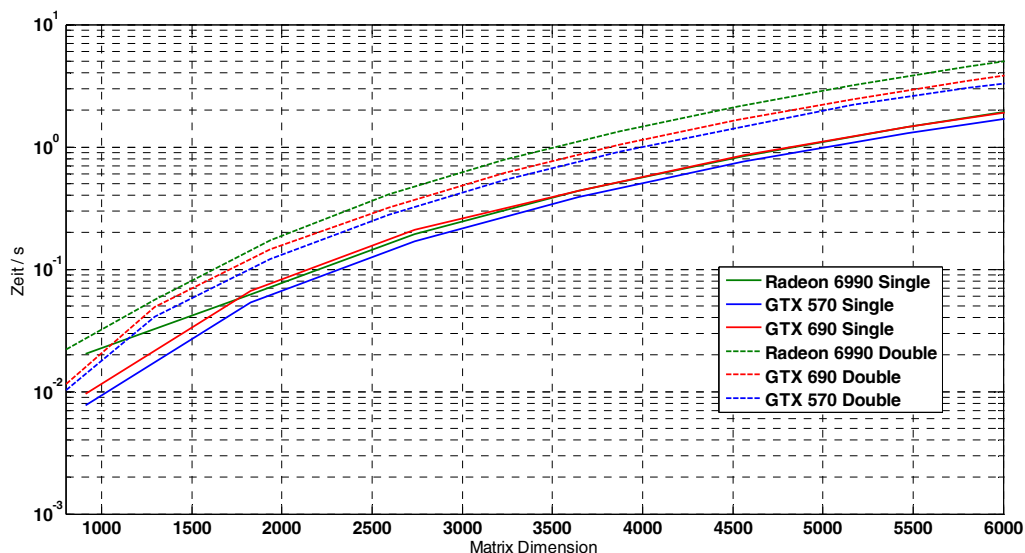


Abbildung 13: Exemplarische Darstellung eines Benchmarks zur optimierten Matrixmultiplikation in OpenCL. Hierbei wird deutlich, dass eine GTX690 für große Datenmengen sogar langsamer sein kann als eine GTX570.

Die direkte naive Portierung erzielte eine Verbesserung der Detektionszeit auf 90ms, durch weitere (teilweise architekturenspezifische) Optimierungen konnte dieser Wert auf 60ms reduziert werden. Diese Werte wurden durch den Einsatz von Radeon 7970 GPUs erreicht. Es wurde somit möglich mit nur einer GPU videotaktschritthaltend Detektionen durchzuführen. Durch den zusätzlichen Zeitpuffer

von 40ms, ist es möglich weitere Verarbeitungsschritte einzufügen ohne die Möglichkeit einer videotaktschritthaltenden Verarbeitung zu verlieren. Erstaunlicherweise zeigte die Verwendung von GPUs der nachfolgenden Generation, d.h. Radeon R9 290x, keinerlei Geschwindigkeitsgewinn. Die Detektionsgeschwindigkeit erhöhte sich sogar auf bis zu 6 Sekunden. Es stellte sich heraus, dass dieser Wert durch einen unausgereiften Treiber hervorgerufen wurde, welcher aufgrund der algorithmischen Struktur des Detektionsverfahrens nicht in der Lage war die enthaltenen Unterroutrinen in adäquater Zeit aufzurufen.

5.2.4 TRAINING DES DETEKTORS MIT KAMERASPEZIFISCHEN MERKMALEN

Zusätzlich zur Optimierung der Verarbeitungsgeschwindigkeit wurden ebenfalls Verbesserungen im Aspekt der Detektionsgüte durchgeführt. Das Detektionsverfahren verwendet primär eine Support Vector Maschine (SVM) um Detektionen hinsichtlich ihrer Güte zu beurteilen. Diese SVM wurde zunächst auf einem generischen Datensatz von Personenbildern trainiert, dadurch ist die SVM in der Lage mit Personenmerkmalen umzugehen, welche eine hohe statistische Varianz auf einem großen Bereich generischer Szenarien besitzen. Diese zunächst positiv klingende Eigenschaft ist jedoch im Detail zu betrachten. Durch den Einsatz fest positionierter Kameras ist es nicht nötig generische Szenarien zu adressieren. Weiterhin ist eine Verbesserung der Detektionsgüte zu erwarten, sofern die SVM auf einem großen Datensatz von kamerabezogenen Aufnahmen trainiert wird. Hierdurch können intrinsische Szenenmerkmale einschließlich ihrer statistischen Varianz berücksichtigt werden.

Unter Benutzung der aufgenommen Sequenzen wurden kameraspezifische Detektoren trainiert, dies resultierte in teilweise sehr deutlichen Verbesserungen. Ein solches Beispiel wird in Abbildung 14 dargestellt, hierbei ist zu erkennen, dass unter Verwendung eines spezifischen Datensatzes wesentlich bessere Ergebnisse erzielt werden.



Abbildung 14: Exemplarischer Vergleich zwischen einem auf generischen Daten trainierten Personendetektors (links) und einem auf kameraspezifischen Daten trainierten Detektors (rechts).

5.2.5 FUSION MULTIPLER DETEKTOREN

Durch die Fusion multipler (körperteil-bezogener) Detektoren D_1, \dots, D_n ist es möglich, einen neuen Detektor D zu kreieren, welcher (körperteil-bezogen) Personen zuverlässiger lokalisieren kann als die Einzeldetektoren [23]. Hierzu werden, wie in Abbildung 15 dargestellt, die Ergebnisse der Einzeldetektoren gruppiert.



Abbildung 15: Kombination von 2 Detektoren: Sinnvoll passende Kopf- und Oberkörper-Detektionen werden gruppiert durch kombinatorische Paarbildung. Nicht gruppierte Detektionen werden verworfen.

Sofern sich die Detektionen hinsichtlich ihrer Größe und Position in einem zulässigen Verhältnis bewegen, so werden diese zu einer neuen Detektion zusammengefasst. Alle übrigen Detektionen, d.h. jene welche nicht mit anderen kombiniert werden können, werden verworfen. Im Projektrahmen wurden Oberkörper- sowie Kopf-Detektoren verwendet um einen zuverlässigen Oberkörper-Detektor zu erhalten. Durch die illustrierte Gruppierung wird die statistische Korrelation von korrekt lokalisierten Körperpartien ausgenutzt, um die Menge von positiven Detektions-Kandidaten zu erhöhen und Falschdetektionen der einzelnen Detektoren zu unterdrücken. Die Resultate dieser Fusion sind in Abbildung 16 dargestellt. Hierbei wird deutlich welche Verbesserung erzielt werden kann (insb. im direkten Vergleich mit Abbildung 14).



Abbildung 16: Detektionsgüte von fusionierten Detektoren (Kopf- und Oberkörper-Detektor).

5.2.6 NICHTLINEARE METRIK FÜR SCHWELLWERTBASIERTE DETEKTIONSAUSWAHL

Ein weiteres Problem welches im Rahmen des Projekts adressiert wurde, ist die schwellwert-basierte Auswahl von SVM Ergebnissen. Um die Menge von falsch-positiven Detektionen gering zu halten werden unter allen Detektionskandidaten nur jene ausgewählt, für welche die SVM eine vorgegebene Güte attestiert (d.h. einen Schwellwert übersteigt). Für im Bild groß erscheinende Personen bzw. Körperpartien liefert die SVM in der Regel geringere Gütewerte als für klein erscheinende. Somit kann es durchaus dazu kommen, dass Personen in bestimmten Bildarealen nicht mehr lokalisiert werden. Ein Bsp. wird im linken Bild in Abbildung 17 gezeigt, hierbei wird der Oberkörper der Person im oberen Bildbereich korrekt lokalisiert, jedoch wird die Person im unteren Bildfeld nicht mehr wahrgenommen. Die Herabsetzung des Schwellwerts führte in der Regel zu vielen falsch-positiven Detektionen, welche mit detektor-spezifischen Methoden nicht mehr effektiv herausgefiltert werden können.

Dieses Problem wurde durch eine Neuskalierung der Gütewerte $val(r)$ erreicht. Hierbei wird eine nichtlineare Skalierung durch eine Funktion f vorgenommen.

$$val_{r_x, r_y, r_s}(r) = f(r_x, r_y, r_s) * val(r)$$

Der Skalierungswert wird abhängig von der Position (x, y) und Größe (s) der Detektion bestimmt.

$$f(r_x, r_y, r_s) = \begin{cases} \tau * \exp(-\beta((I_h - r_y)/I_h)) * r_s, & r_s \geq \omega \\ 1, & r_s < \omega \end{cases}$$

Die Bedeutung der Parameter ist wie folgt: ω ist der Schwellwert für die Größe der Detektion, I_h ist die Bildhöhe und β, τ Skalierungsfaktoren. Das Prinzip wird im mittleren Bild von Abbildung 17 veranschaulicht, je nichttransparenter bzw. dunkler das Bild wird umso mehr nähert sich der Skalierungsfaktor dem Wert 1 an, d.h. es werden nur Detektion im unteren Bildbereich skaliert. Hierbei wird jedoch zusätzlich unterschieden welche Größe die Detektion hat, d.h. kleine Detektionen werden nicht skaliert. Wie im rechten Bild in Abbildung 17 exemplarisch dargestellt wird, führte dieses Vorgehen zur gewünschten Verbesserung. Die Person im unteren Bildbereich wird korrekt detektiert wohingegen die obere Detektion unbeeinflusst bleibt.



Abbildung 17: Skalierung der Güterwerte einer SVM: Links: Beispiel für naive schwellwert-basierte Selektion von Detektionen. Mitte: Über das Bild gelegte Skalierungsfunktion, nichttransparente/dunkle Flächen indizieren einen Skalierungswert nahe 1, transparente Flächen einen Wert größer 1, Rechts: Resultat bei nichtlinearer Skalierung der Güterwerte einer SVM

5.2.7 PERSONENDETEKTION IN PARALLELEN VIDEODATENSTRÖMEN

In der praktischen Anwendung kommen in der Regel mehrere Kameraströme vor. Dabei ist es möglich jeden Strom durch einen separaten dezentralen Rechenknoten zu verarbeiten. Zur Bewältigung des anfallenden Videomaterials können aber durchaus auch mehrere GPUs in einem einzelnen Computer zur Verfügung gestellt werden. Jedoch liegt dabei die Obergrenze meist bei 4 Dual-Grafikkarten (2 GPUs pro Karte), d.h. der Verarbeitung von 8 Videoströmen. Da jedoch nicht jeder Rechner über eine solche Ausstattung verfügt und in der Regel mehr als 8 Videoströme anfallen wurde eine Architektur entwickelt, welche es erlaubt das Videomaterial über Netzwerkverbindungen auf mehrere unabhängige Rechensysteme mit multiplen Konfigurationen zu verteilen [24] [25]. Die Systemstruktur ist in Abbildung 18 dargestellt.

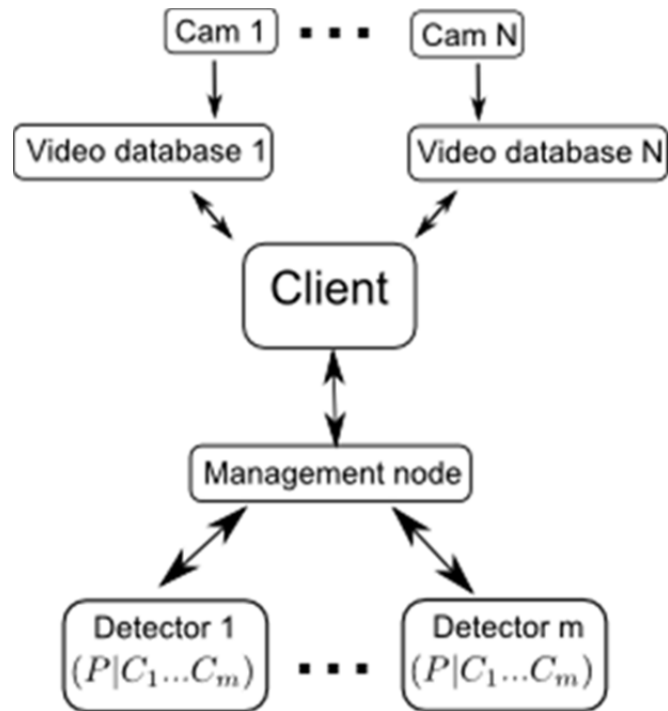


Abbildung 18: Cluster: Die Abb. zeigt die Struktur des Clustersystems zur Verarbeitung paralleler Videoströme.

Der Client erhält hierbei das gewünschte Videomaterial direkt aus der Videodatenbank. Sofern für darin enthaltene Bilder eine Detektion gewünscht ist, wird die Bildmenge an einen Managementknoten geschickt. Dieser bereitet die Anfrage für eine effiziente Verarbeitung auf und delegiert die anfallenden Detektionsvorgänge an die im Netzwerk verfügbaren Detektoren. Der konkret umgesetzte Rechencluster besteht aus 11 Knoten, 8 davon ausgestattet mit hochperformanten Grafikkarten von AMD, dadurch war es möglich mindestens 8 parallel eintreffende Videoströme zu verarbeiten. Ein weiterer Vorteil dieser Systemstruktur ist die einfache Erweiterbarkeit sofern mehr Rechenkapazität benötigt wird, es ist jederzeit möglich weitere Knoten in das System zu integrieren. Somit bietet die entwickelte Struktur eine mit der Knotenzahl linearskalierende Kapazität sowie die Möglichkeit zur schnellen Adaption an die Eigenschaften des Einsatzgebiets.

Zur weiteren Steigerung der Effizienz wurden auf der Seite der Detektoren redundante Berechnungen größtenteils eliminiert. Dies wird ebenfalls in Abbildung 18 illustriert, jeder Detektor führt für ein neues Bild die Vorverarbeitung ‚P‘ nur einmal aus, anschließende Detektionsprozesse ‚C‘ können auf diese vorverarbeiteten Daten zugreifen. Insgesamt lassen sich dadurch auf einem einzelnen Detektor $(n-1) \cdot 40\text{ms}$ Berechnungszeit für multiple Detektionen auf einem Bild einsparen, hierbei symbolisiert n die Anzahl der durchzuführenden Detektionen. Hierdurch wurde es möglich auf einer GPU bis zu 3 Detektionen pro 100ms durchzuführen, d.h. eine Verdreifachung der Effizienz wurde erreicht. Dies ist insbesondere für die zuvor beschriebene Nutzung von multiplen Detektoren nützlich, auf diesem Weg kann eine Steigerung der Detektionsgüte bei videotaktischhaltender Verarbeitung erreicht werden.

5.2.8 EVALUATION NICHTLINEARER KERNEL

Support Vector Maschinen entscheiden über die Güte einer Detektion indem sie sogenannte Kernel verwenden. Ein Kernel ist eine mathematische Funktion und dient der Bestimmung eines Ähnlichkeitswertes für hochdimensionale Vektoren. Personenbilder werden in diesem Kontext als

Vektoren mit mehr als 3000 Elementen dargestellt. Oftmals werden lineare Kernel verwendet, hierbei handelt es sich z.B. um das kanonische Skalarprodukt. Es ist jedoch möglich nichtlineare Kernel zu verwenden, diese bieten in der Regel eine höhere Genauigkeit bei der Klassifikation von Daten. Ein Beispiel für solche Funktionen ist die Gaußfunktion $\exp(-\|x-y\|)$, welche zur Gruppe der radialen Basisfunktionen gehört. Im Rahmen des Projekts wurde die Nützlichkeit dieser Funktionen bezogen auf stark variierende Hintergründe untersucht. Ein Anwendungsbeispiel ist die Situation in welcher eine Person einer Personengruppe beitrifft und in dieser temporär partiell verdeckt wird. Ein klassischer HOG Detektor versagt in solchen Fällen häufig.

Ein entscheidender Nachteil der nichtlinearen Funktionen ist deren Auswirkung auf die Berechnungskomplexität innerhalb der SVM. Während bei einem linearen Kernel die Trainingsdauer für eine SVM 30 Minuten betragen kann, so erhöht sich dies bei Verwendung nichtlinearer Funktionen auf 20-24 Stunden. Auch im Hinblick auf die Detektionszeit ergeben sich große zeitliche Nachteile, die Detektionszeit wächst selbst bei Verwendung einer GPU stark an. Während im linearen Fall nur 60ms nötig sind, so sind hierbei 35s nötig.

Eine Analyse der Detektionsgüte zeigte auf, dass die Nutzung von radialen Basisfunktionen keinen Gewinn brachte. Zusätzlich wurde beobachtet, dass während des Trainings einer SVM oftmals das Problem des Overfittings auftrat, d.h. die SVM lernte auf dem Trainingsdatensatz auswendig anstatt die generischen Merkmale der Daten zu nutzen.

5.2.9 KAMERABEZOGENE ERGEBNISFUSION

In diesem Projekt kamen viele unterschiedliche Bildverarbeitungsverfahren zum Einsatz, welche das Ziel verfolgen spezielle Bildregionen zu detektieren, wie z.B. Bewegungsdetektoren, Vordergrundsegmentierung, Personen-, Kopf-, Kopf-/Schulter-Detektoren und Gesichtsdetektoren. Eine sehr grundlegende Rolle im APfel-Projekt spielt unter anderem die Ganzkörperdetektion von Personen. Je nach Szenario kann es hierbei zu einer großen Anzahl an Falsch-Positiv-Detektionen kommen (Abbildung 19).

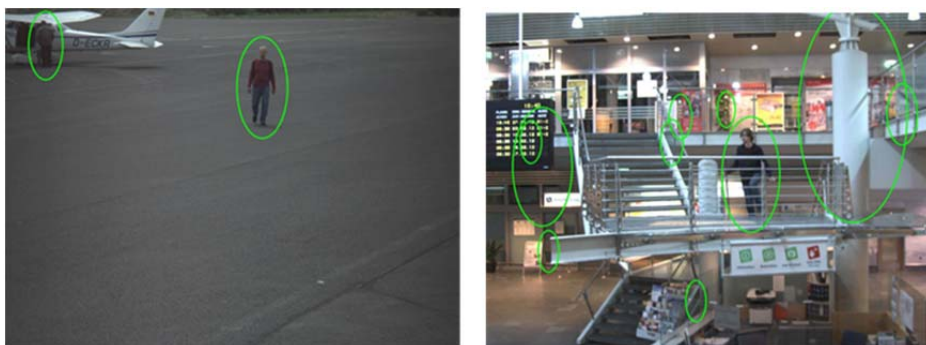


Abbildung 19: Beispielergebnisse eines Detektionsverfahrens (l.: sehr gute Detektionsleistung bei homogenem Hintergrund; r.: viele Falsch-Positiv-Detektionen).

Um dieses Problem zu lösen, wurde von der HRW ein Fusionsmodul entwickelt [22], welches die Ergebnisse der unterschiedlichen Verfahren miteinander kombiniert. Die Ergebnisse (in Form von ROIs) der einzelnen Verfahren werden hierfür zunächst von dem Fusionsmodul aus der zentralen Datenbank ausgelesen. Auf diese Weise ist eine sehr einfache, partnerübergreifende Integration der Verfahren möglich. Ausgehend von einer Personendetektion wird in einem nächsten Schritt nach Überschneidungen mit den ROIs bestimmter Verfahren gesucht. Anhand dieser Überschneidungen wird ein Score berechnet und die ROI der Personendetektion als positiv klassifiziert, wenn der Score

größer als ein festgelegter Schwellwert ist. Sollte beispielsweise eine Personendetektion vorliegen, in der keine Bewegung detektiert wurde, und die auch nicht Teil des Vordergrundes ist, so ist die Wahrscheinlichkeit, dass es sich hierbei um eine Fehldetektion handelt, sehr hoch. In einem solchen Fall wird die ROI der Personendetektion verworfen. Der Fusionsprozess wird in Abbildung 20 veranschaulicht.

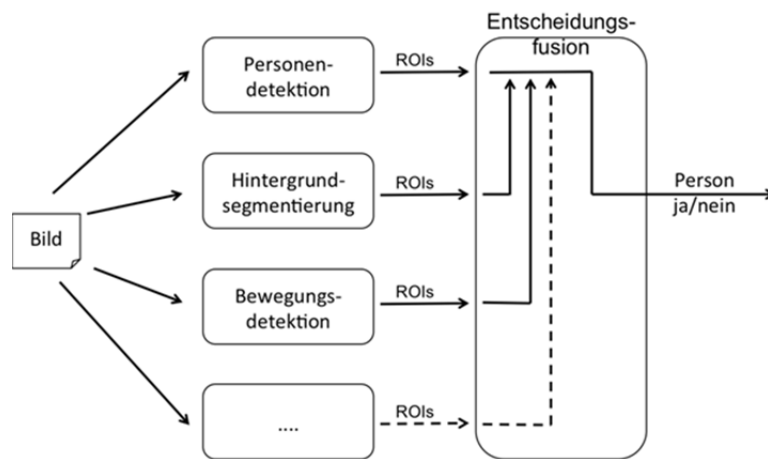


Abbildung 20: Fusionsprozess. Die Ergebnisse der verschiedenen Module werden kombiniert. In Abhängigkeit eines Scores entscheidet der Prozess ob eine Personendetektion verworfen oder als positiv klassifiziert wird.

Es ergeben sich hieraus verschiedene Kombinationsmöglichkeiten. So lassen sich z. B. Kopf und Ganzkörperdetektion miteinander fusionieren. Eine Personendetektion gilt in dem Fall nur dann als positiv, wenn ein Kopf im oberen Viertel detektiert wurde.

Durch das entwickelte Fusionsmodul wird das Ergebnis gegenüber den Einzelmodulen erheblich verbessert, da dem Entscheidungsprozess mehr unabhängige Informationen vorliegen. Der Einsatz des Fusionsmoduls hat gezeigt, dass Personen und Köpfe auf diese Weise sehr robust detektiert werden können. Die Falschdetektionen werden erheblich reduziert, was zu einer Erhöhung der Zuverlässigkeit des Gesamtsystems führt. Ein Beispiel ist in Abbildung 21 zu sehen.

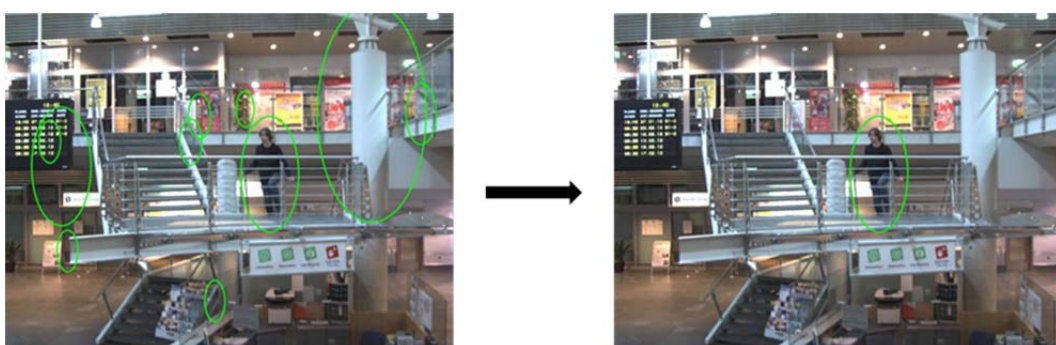


Abbildung 21: Ergebnis des Fusionsprozesses. Durch die Kombination verschiedener Module wurden die Falschdetektionen (l.) um 100 % reduziert (r.).

Die Anwendung dieses Verfahrens zeigt das Potenzial von Fusionsverfahren. Die beschriebene Methode ist jedoch davon abhängig, dass ein möglichst robuster Personendetektor eingesetzt wird, der sämtliche Personen auf einem Bild findet. Schlägt der Detektor nicht an, findet keine Fusion statt. Das Verfahren wurde deshalb dahingehend weiterentwickelt, dass auch diejenigen Personen gefunden werden, die der Detektor nicht ermitteln konnte. Um dies zu erreichen, wird in einem nächsten Schritt jedes Verfahren unterschiedlich gewichtet. Eine geeignete Gewichtung wird dabei

durch eine Verfahrensevaluation ermittelt. Die Erweiterung des Fusionsprozesses ist in Abbildung 22 dargestellt. Jeder Pixel einer ROI wird entsprechend gewichtet, wobei die Gewichte übereinanderliegender ROIs aufsummiert werden. Markiert werden dann diejenigen Bereiche, die über einem spezifizierten Schwellwert liegen. Das Ergebnis ist eine Aufmerksamkeitskarte (rechts im Bild). Durch diese Erweiterung wird eine Unabhängigkeit von der Personendetektion erreicht bei gleichzeitiger Eliminierung von Falsch-Positiven-Resultaten.

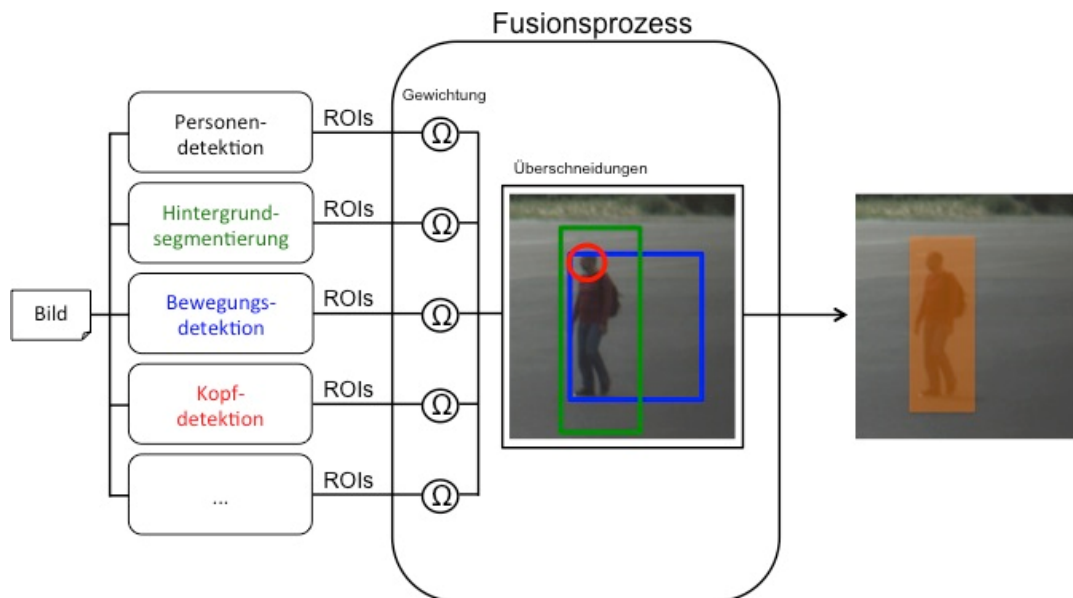


Abbildung 22: Gewichteter Fusionsprozess. Durch die Aufsummierung der einzelnen Verfahren wird die Person im Bild markiert, obwohl der Personendetektor selbst keine ROI liefert.

5.3 WIEDERERKENNUNG

Im Verlauf dieses Projektes lag der Arbeitsschwerpunkt der HRW im Bereich der Wiedererkennung zunächst auf der kamerabezogenen Wiedererkennung. In der einjährigen Verlängerungsphase wurden diese Verfahren auf die Kopf-Schulter-Partie übertragen und für eine kameraübergreifende Wiedererkennung erweitert.

5.3.1 KAMERABEZOGENE WIEDERERKENNUNG

Die kamerabezogene Wiedererkennung erlaubt die Entscheidung darüber, ob sich eine Person noch im aktuellen Überwachungsbereich befindet, oder ob eine kameraübergreifende Wiedererkennung angestoßen werden muss. Hierzu werden die durch den Merkmalstracker extrahierten Merkmale verwendet. Diese Tracksegmente können anschließend logisch zu einem kompletten Track einer Person verknüpft werden (TrackletToTrack). Aus diesem Track kann nun ein Merkmalsraum der Person gebildet werden. In einer ersten Version wurden nach dem Markieren einer Person und Festlegen des zeitlichen und räumlichen Suchraums, durch die von Projektpartnern zur Verfügung gestellten Reasoningkomponente, alle extrahierten Merkmale aus dem Suchraum angefragt und mit den Merkmalen des Merkmalsraumes verglichen. Nach Sortierung führt dies zu einem Hypothesenranking aller sich im Suchraum befindlichen Personen. Die finale Version dieses Verfahrens wurde allerdings in die dezentrale Liveanalyse eingebettet, so dass permanent zu jedem Merkmalstrack neue Kandidaten gesucht und gegebenenfalls zu einem Track zusammengefasst werden. Um in einem Dauerbetrieb nicht unendlich viele Merkmalstracks betrachten zu müssen, wird hier eine maximale zeitliche Hysterese festgelegt, welche je nach Einsatzzweck variieren kann. In diesem Projekt werden dabei Merkmalstracks nur bis zu 10s in der Vergangenheit betrachtet,

wenn bis dahin nicht neu verknüpft wurde. Die kamerabezogene Wiedererkennung ist exemplarisch in Abbildung 23 dargestellt.

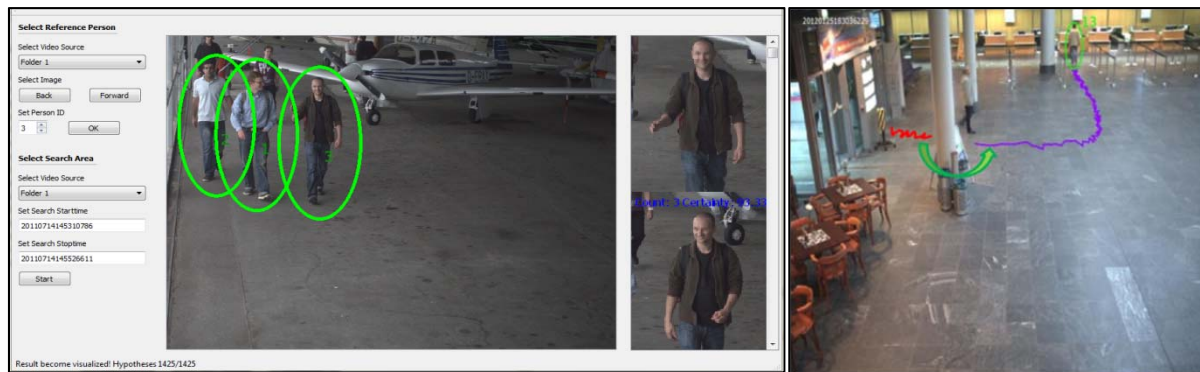


Abbildung 23: Kamerabezogene Wiedererkennung: Im linken Bild wurde exemplarisch Person 3 markiert (mitte) und mit einer hohen Übereinstimmung aus 1425 Hypothesen wiedererkannt (rechts). Im rechten Bild ist die Vereinigung zweier Tracklets zu einem Track visualisiert.

5.3.2 KAMERAÜBERGREIFENDE WIEDERERKENNUNG

Während der einjährigen Verlängerungsphase übernahm die Hochschule Ruhr West die Entwicklung eines weiteren Wiedererkennungsmoduls. Dieses soll Personen auch bei Verdeckung innerhalb einer Gruppe wiedererkennen. Hierzu betrachtet die HRW den Kopf-Schulter-Bereich einer Person. Innerhalb dieses Detektionsfensters wurde ein fixer Bereich ausgewählt, aus welchem die zur Wiedererkennung verwendeten Merkmale extrahiert werden (Abbildung 24).

Um eine kameraübergreifende Wiedererkennung zu ermöglichen, wurde der bereits zur kamerabezogenen Wiedererkennung verwendete Merkmalsatz (allgemeine Merkmale) um individuelle Merkmale erweitert. Im Gegensatz zu den allgemeinen Merkmalen variieren die individuellen Merkmale von Person zu Person und zwischen verschiedenen Ansichten in ihrer Position, Anzahl und Ausprägung. Diese individuellen Merkmale sollen Auffälligkeiten der Kleidung repräsentieren, eine sich von der Umgebung abhebende Farbe oder Textur welche sehr wahrscheinlich diskriminierend für die Person ist, in den allgemeinen Merkmalen aber nicht hinreichend repräsentiert wird. Um die Varianz dieser Merkmale aufgrund der Ansicht zu berücksichtigen, werden die allgemeinen Merkmale weiterhin videotaktschritthaltend getrackt und mittels dem oben beschriebenen TrackletToTrack verknüpft. Dadurch erhält man lange Tracks, welche verschiedene Ansichten enthalten.

Zur Lokalisation der individuellen Merkmale kommt eine adaptierte und erweiterte Implementierung des Verfahrens von Itti und Koch [26] zum Einsatz [23]. Hierbei wird zum einen der HSV-Raum verwendet zum anderen werden die einzelnen Aufmerksamkeitskarten, für die Farbe, Textur und Intensität, nicht weiter vereinigt. Die Nutzung der Farbwerte (hue) des HSV-Raums hat einen ähnlichen Effekt wie die Kombination des RGB-Raumes zu einem Rot-Grün- und einem Blau-Gelb-Kanal. Bei der Nutzung des losen Farbwertes bleiben allerdings die Intensität und die Helligkeit unberührt. Im Originalverfahren wird als Intensität die Helligkeit im RGB-Raum, also das Grauwertbild betrachtet, wogegen wir die Farbsättigung (saturation) aus dem HSV-Raum nutzen. Die Orientierung wurde dagegen genau wie im Originalverfahren bestimmt. Zusätzlich wurde das Verfahren um eine Aufmerksamkeitskarte erweitert. Hierzu wurde die lokale Entropie [27] [28] auf neun Skalierungsstufen des Grauwertbildes bestimmt und äquivalent zu den anderen zu einer Aufmerksamkeitskarte kombiniert. Diese zusätzliche Karte gibt an, an welchen Stellen im

ausgewählten Bereich der Detektion starke Schwankungen in der Helligkeitsverteilung vorliegen. Abbildung 24 zeigt schematisch das adaptierte Verfahren.

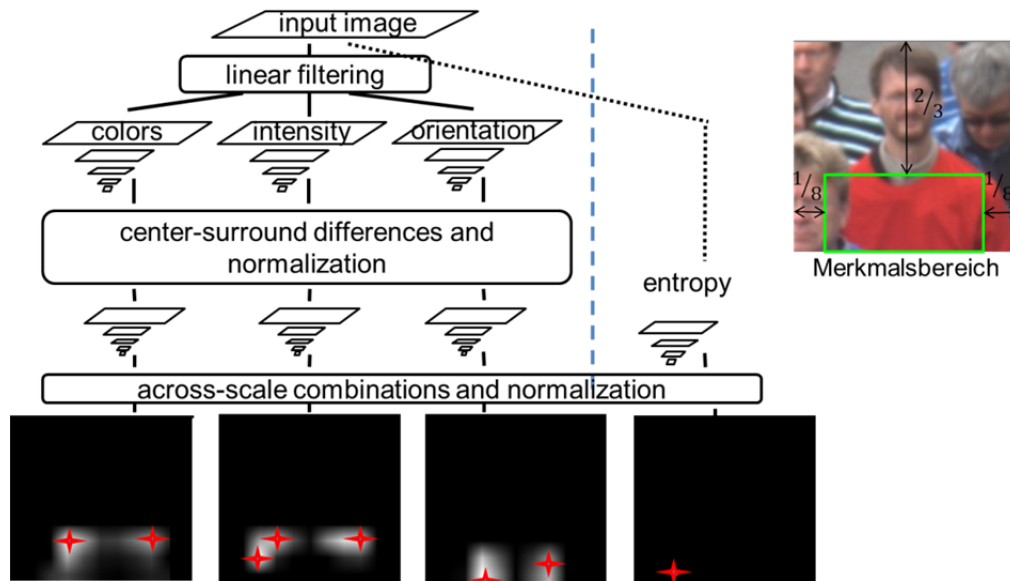


Abbildung 24: individuelle Merkmale: Mit dem hier dargestellten Verfahren werden auffällige Bereiche innerhalb des Merkmalsbereiches lokalisiert. Der Merkmalsbereich ist innerhalb der Detektion fix lokalisiert.

Da die Berechnung der individuellen Merkmale sehr rechenintensiv ist wurde darauf verzichtet diese bereits videotaktschritthaltend für jede Detektion zu bestimmen. Vielmehr kommt ein kaskadierender Merkmalsvergleich zum Einsatz, der es ermöglicht die individuellen Merkmale für nur wenige Detektionen zu bestimmen. Hierbei werden zunächst zum Zeitpunkt der Wiedererkennung alle Detektionen der Tracks aus dem Suchraum mit den Detektionen des Tracks der gesuchten Person mittels der allgemeinen Merkmale verglichen. Aus jedem Track des Suchraumes wird die Detektion mit der korrespondierenden Detektion der gesuchten Person verknüpft, welche den geringsten Fehler zueinander aufweisen. Zusätzlich wird ein Schwellwert für diesen Fehler bestimmt. Liegt der ermittelte Fehlerwert eines Detektionspaares über dem Schwellwert wird dieses Paar und damit der Track nicht weiter betrachtet, da davon auszugehen ist, dass dies nicht dieselben Personen sein können. Befindet sich die gesuchte Person unter den übrigen Detektionen ist die Wahrscheinlichkeit sehr hoch, dass das Detektionspaar eine ähnliche Ansicht beider Detektionen aufweist. Daher werden im Folgenden die individuellen Merkmale nur für die Detektionspaare bestimmt. Dieses Vorgehen ist in Abbildung 25 dargestellt.

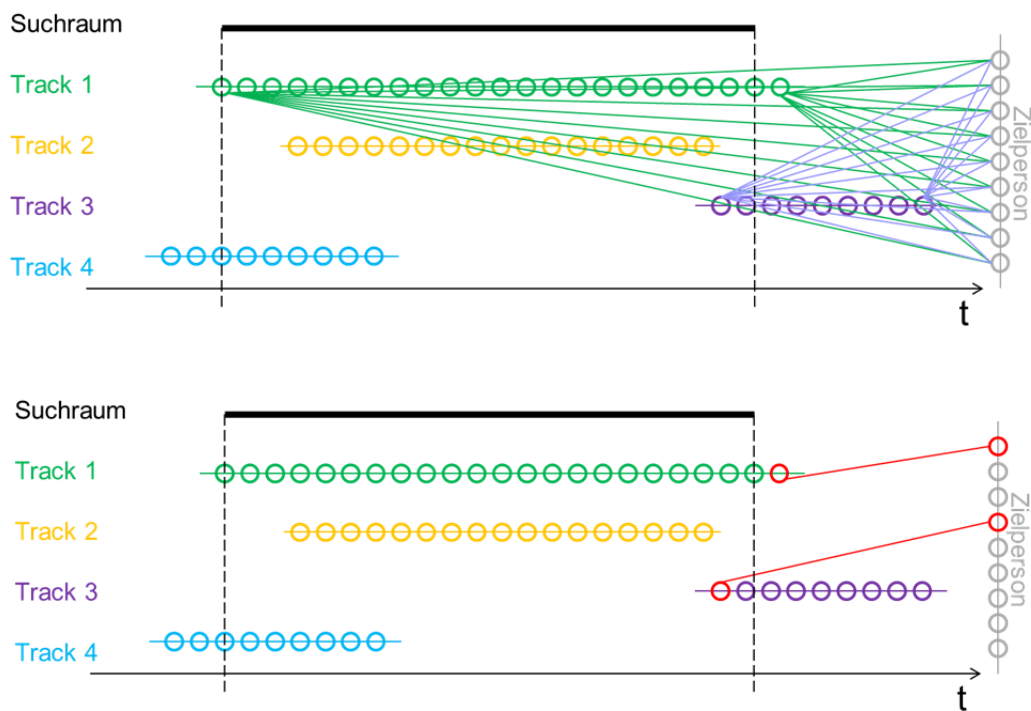


Abbildung 25: kaskadierender Merkmalsvergleich: oben: die allgemeinen Merkmale werden zwischen allen Detektionen verglichen (hier zur Übersichtlichkeit nicht vollvermascht dargestellt); unten: die individuellen Merkmale werden nur für wenige Detektionspaare bestimmt und verglichen

Beim Vergleich der individuellen Merkmale werden zunächst alle Aufmerksamkeitskarten separat betrachtet. Zum einen wird hierbei der Abstand zwischen den Merkmalen beider Detektionen betrachtet (normalisierte euklidische Distanz), und zum anderen werden die Werte an den lokal korrespondierenden Merkmalen verglichen (normalisierte absolute Differenz). Diese Fehlerwerte werden für jede Aufmerksamkeitskarte summiert und durch ihre Anzahl geteilt. Für alle Aufmerksamkeitskarten werden die Fehler ebenfalls aufsummiert und durch 4 (Anzahl der Karten) geteilt. Eine Gewichtung der Karten war dabei nicht zielführend. Der Fehler der individuellen Merkmale wird mit 0,5 gewichtet und anschließend auf den Fehler der allgemeinen Merkmale addiert. Diese Summe wird durch 1,5 geteilt, damit der resultierende Fehler wieder zwischen 0 und 1 liegt.

5.3.2.1 Evaluation der kameraübergreifenden Wiedererkennung

Zur Bewertung der kameraübergreifenden Wiedererkennung wurden die am Flughafen Erfurt-Weimar aufgezeichneten Sequenzen verwendet. Als elementares Wiedererkennungsmerkmal wurde der Kopf-Schulter-Bereich einzelner Personen gewählt (siehe Abbildung 26). Im mittleren sowie im rechten Bild von Abbildung 26, welche jeweils aus Szenen am Flughafen Erfurt-Weimar stammen, wird bereits deutlich, welche Faktoren zusätzlich zur Gruppenproblematik die Personenerkennung erschweren. Personen können mit verschiedenen Auflösungen abgebildet sein, hierbei sind besonders geringe Auflösungen (Abbildung 26, rechtes Bild) eine Herausforderung da nahezu keine deskriptiven Merkmale extrahiert werden können. Personen mit ähnlicher Kleidung sind dadurch teilweise nicht mehr unterscheidbar. Zusätzlich ergeben sich durch lange Belichtungszeiten im Innenbereich Bewegungsunschärfe (Abbildung 26, mittleres Bild), was zu einer Detailreduktion in extrahierten Merkmalen führt.

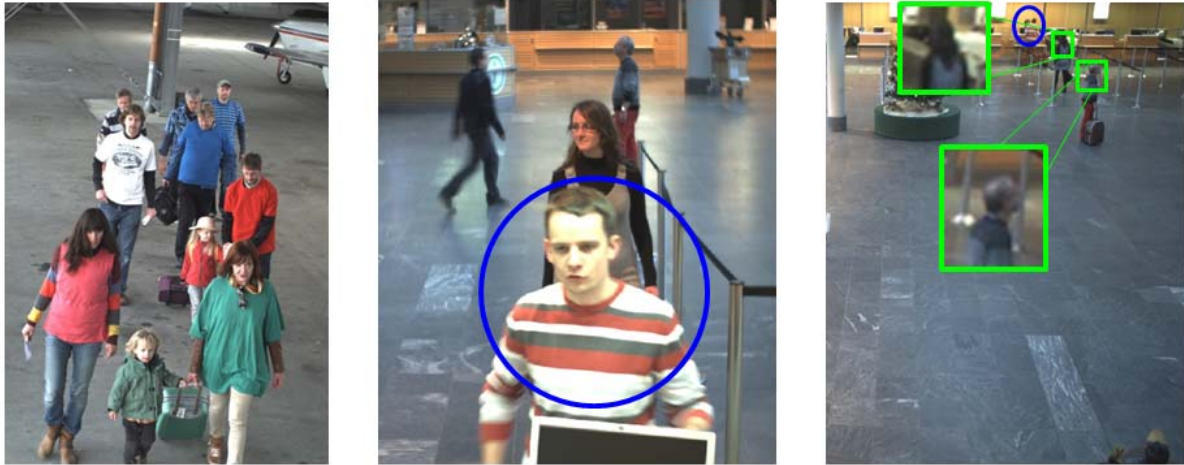


Abbildung 26 Exemplarische Bilder aus beiden Sequenzen. Links: Schönhagen Kamera C7; mitte: Erfurt-Weimar Kamera C1; rechts: Erfurt-Weimar Kamera C2. Im linken Bild sind alle Personen in der Gruppe klar dargestellt während im mittleren Bild eine deutliche Bewegungsunschärfe erkennbar ist. Die Personen im rechten Bild sind aufgrund Szenentiefe sehr klein dargestellt, der Kopf-Schulter-Bereich zweier Personen ist vergrößert hervorgehoben. Blaue Kreise indizieren die zu erkennende Person auf verschiedenen Kameras.

In Abbildung 27 wird die Falschakzeptanzrate (FAR) dem Detektionsfehler (Error) von Merkmalen gegenübergestellt. Hierbei beinhaltet der Detektionsfehler Falsch-Positive Detektionen als auch

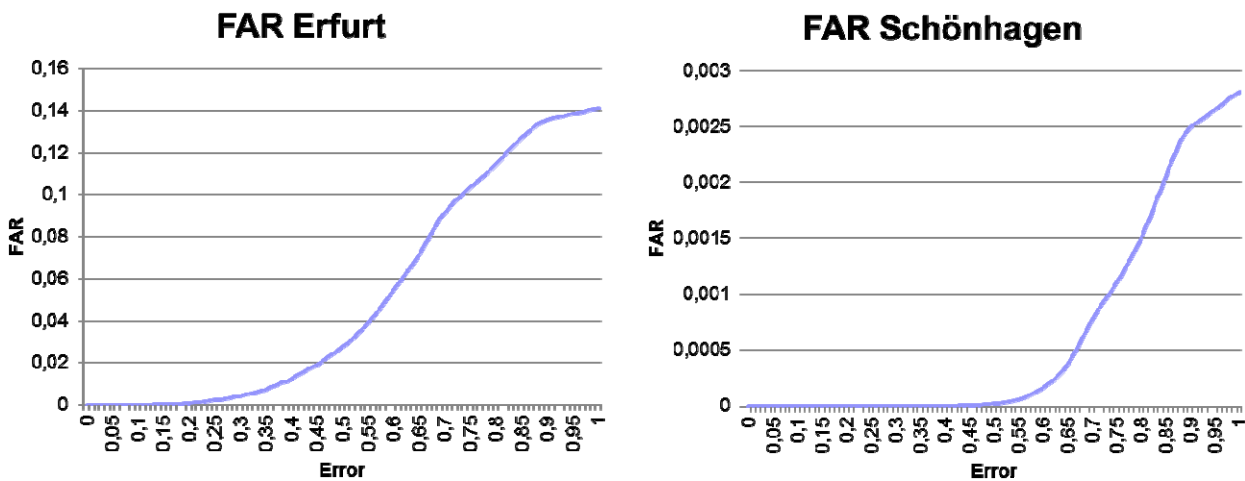


Abbildung 27 Darstellung der Falschakzeptanzrate (FAR) für beide Sequenzen (links: Flughafen Erfurt-Weimar; rechts: Flugplatz Schönhagen) in Abhängigkeit des Detektionsfehlers von Merkmalen.

ungenau lokalisierte korrekte Detektionen. Es ist deutlich zu erkennen, dass im Hinblick auf die Sequenzen vom Flugplatz Schönhagen, fehlerhafte Detektionen erst bei einem relativ hohen Fehlerwert zu einem Anstieg der FAR führen. Weiterhin ist der maximale FAR-Wert wesentlich geringer als bei Sequenzen vom Flughafen Erfurt-Weimar. Diese Ergebnisse korrelieren stark mit den zuvor erläuterten Beobachtungen hinsichtlich Abbildung 26. Unter Verwendung dieser Resultate kann gefolgert werden, dass in Bezug auf Aufnahmen am Flugplatz Schönhagen auch kleinere Merkmale wie z.B. nur der Brustbereich bereits adäquate Wiedererkennungsraten liefern können. Hierbei muss beachtet werden, dass diese Folgerung nur haltbar ist sofern die aufgezeichneten Personen eine vergleichbare Merkmalsvarianz zu jenen in den verwendeten Evaluationsaufnahmen aufweisen. Die Robustheit des kameraübergreifenden Erkennungsverfahrens wird jedoch bei Betrachtung der maximalen FAR bzgl. der Erfurt-Weimar Sequenzen deutlich, selbst bei maximalen Detektionsfehlern und teilweise problematischen Aufnahmen liegt die FAR bei lediglich 14%.

Weiterhin zeigen die Ergebnisse die Abhängigkeit von FAR und einer robusten Merkmalsdetektion auf, ein zuverlässiger Detektor kann hierbei ungünstige Aufnahmeverhältnisse durchaus kompensieren.

5.3.3 EVALUATION ALLER VIDEOINDIZIERENDEN VERFAHREN

Ein weiteres Arbeitspaket, für welches die HRW verantwortlich war, ist die Evaluation der videoindizierenden Verfahren, also aller Verfahren welche bereits zum Zeitpunkt der Aufnahme höherwertige Informationen zu den Bildern extrahieren.

Ziel dieser Evaluation ist es, die Qualität aller videoindizierenden Verfahren aller Partner auf ihre Güte und Effizienz unter verschiedenen realen Bedingungen zu evaluieren. Dabei wurde die Analyse speziell auf die in APFeI adressierte Problemstellung angepasst. Zu den zu evaluierenden Verfahren zählen:

- Vordergrundsegmentierung (adaptives Differenzbildverfahren) – HRW
- Bewegungsdetektion – HRW
- GPU-basierte Personendetektion – HRW
- Merkmalstracker – HRW
- Vordergrundsegmentierung (MoG) – TU-Ilmenau
- Grundflächenbasierte Personendetektion – TU-Ilmenau
- Personentracker – TU-Ilmenau
- Gesichtsdetektor und Tracker – L-1

Hierzu wurden zunächst bei den Endanwendern dem Flugplatz Schönhagen und dem Flughafen Erfurt-Weimar, mittels Probanden Testaufnahmen aufgezeichnet. Diese wurden anschließend gelabelt, das heißt alle im Bild enthaltenen Personen, Köpfe und Gesichter wurden händisch markiert. Schließlich wurden alle Verfahren auf die Testsequenzen angewendet und deren Ergebnisse in ein XML-Format überführt. Als Indikator für die Güte der Verfahren wurde die CLEAR-Metrik gewählt. Die normierte „Multiple Object Detection Accuracy (N-MODA)“ gibt einen Wert für die Fehlerrate eines Detektionsverfahrens an, wobei die falsch positiven Detektionen (fp) und die falsch negativen Detektionen (fn) über alle Bilder einer Sequenz (N_{frames}) betrachtet werden. Hierbei kann eine problemspezifische Gewichtung der Fehler erfolgen (c_{fn} für alle fn und c_{fp} für alle fp). Normiert wird dieser Fehlerwert mit der Anzahl aller tatsächlich im Bild enthaltenen Objekte (N_G). Die „Multiple Object Detection Precision (MODP)“ gibt an wie gut die korrekten Detektionen (N_{mapped}) die Objekte lokalisieren. Hierzu wird die Überlappungsrate der tatsächlichen Region (G) mit der detektierten Region (D) bestimmt. Die normierte „Multiple Object Tracking Accuracy (MOTA)“ stellt ähnlich wie die MODA einen Wert dar, der eine problemspezifische Aussage über die Güte eines Trackingalgorithmus zulässt. Zu den bekannten möglichen Fehlern kann es bei diesen Verfahren noch zu geteilten Tracks (*divided*) sowie zu falsch zugeordneten Tracks (*ID-switches*) führen. Diese können ebenfalls problemspezifisch gewichtet werden ($c_{divided}$ für *divided* und c_s für *ID-switches*). Äquivalent zu der Detektionsauswertung kann auch hier eine „Multiple Objekt Tracking Precision (MOTP)“ bestimmt werden. Diese Metrik soll unter verschiedenen Mindestgenauigkeiten bestimmt werden, so dass zum einen eine Aussage über die Fehleranfälligkeit eines Verfahrens, sowie zur Genauigkeit der gefundenen Bildregionen gemacht werden kann.

$$N-MODA = 1 - \frac{\sum_{t=1}^{N_{frames}} (c_{fn}(fn_t) + c_{fp}(fp_t))}{\sum_{t=1}^{N_{frames}} N_G^{(t)}}$$

$$N-MODP = \frac{\sum_{t=1}^{N_{frames}} \frac{\sum_{i=1}^{N_{mapped}^{(t)}} \frac{|G_i^{(t)} \cap M_i^{(t)}|}{|G_i^{(t)} \cup M_i^{(t)}|}}{N_{mapped}^{(t)}}}{N_{frames}}$$

$$MOTA = 1 - \frac{\sum_{t=1}^{N_{frames}} (c_{fn}(fn_t) + c_{fp}(fp_t) + c_d(divided_t) + c_s(ID-switches_t))}{\sum_{t=1}^{N_{frames}} N_G^{(t)}}$$

$$MOTP = \frac{\sum_{i=1}^{N_{mapped}} \sum_{t=1}^{N_{frames}^{(t)}} \left(\frac{|G_i^{(t)} \cap D_i^{(t)}|}{|G_i^{(t)} \cup D_i^{(t)}|} \right)}{\sum_{t=1}^{N_{frames}} N_{mapped}^{(t)}}$$

Hierzu wurde ein Tool entwickelt, welches die zur Auswertung benötigten Daten bestimmt und diese visualisiert. Zudem ist ein Export in ein CSV-Format möglich, was den Import der Daten in weitere Analyseprogramme wie Excel oder Matlab ermöglicht.

Für die Verfahrensevaluation ist es notwendig, verschiedene Sequenzen auszuwählen und diese verfahrensspezifisch nach Schwierigkeit zu kategorisieren. Es wurden insgesamt 7 Sequenzen ausgewählt und gelabelt, so dass für jedes Verfahren verschiedene Schwierigkeitsstufen ausgewählt werden können. Die nachfolgende Abbildung 28 gibt einen Überblick über die Einstufung der Verfahren bezogen auf die jeweilige Sequenz.




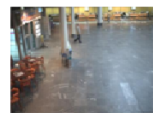
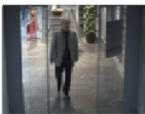


	<table border="1"> <thead> <tr> <th>Verfahren</th> <th>leicht</th> <th>mittel</th> <th>schwer</th> </tr> </thead> <tbody> <tr><td>Vordergrundsegmentierung</td><td>X</td><td></td><td></td></tr> <tr><td>Bewegungsanalyse</td><td>X</td><td></td><td></td></tr> <tr><td>Personendetektion</td><td>X</td><td></td><td></td></tr> <tr><td>Tracker</td><td>X</td><td></td><td></td></tr> <tr><td>Gesichtsdetektion & -tracking</td><td></td><td>X</td><td></td></tr> <tr><td>Kopf-/Kopfschulterdetektion</td><td>X</td><td></td><td></td></tr> </tbody> </table>	Verfahren	leicht	mittel	schwer	Vordergrundsegmentierung	X			Bewegungsanalyse	X			Personendetektion	X			Tracker	X			Gesichtsdetektion & -tracking		X		Kopf-/Kopfschulterdetektion	X				<table border="1"> <thead> <tr> <th>Verfahren</th> <th>leicht</th> <th>mittel</th> <th>schwer</th> </tr> </thead> <tbody> <tr><td>Vordergrundsegmentierung</td><td></td><td></td><td>X</td></tr> <tr><td>Bewegungsanalyse</td><td>X</td><td></td><td></td></tr> <tr><td>Personendetektion</td><td>X</td><td></td><td></td></tr> <tr><td>Tracker</td><td></td><td></td><td>X</td></tr> <tr><td>Gesichtsdetektion & -tracking</td><td>X</td><td></td><td></td></tr> <tr><td>Kopf-/Kopfschulterdetektion</td><td>X</td><td></td><td></td></tr> </tbody> </table>	Verfahren	leicht	mittel	schwer	Vordergrundsegmentierung			X	Bewegungsanalyse	X			Personendetektion	X			Tracker			X	Gesichtsdetektion & -tracking	X			Kopf-/Kopfschulterdetektion	X		
Verfahren	leicht	mittel	schwer																																																								
Vordergrundsegmentierung	X																																																										
Bewegungsanalyse	X																																																										
Personendetektion	X																																																										
Tracker	X																																																										
Gesichtsdetektion & -tracking		X																																																									
Kopf-/Kopfschulterdetektion	X																																																										
Verfahren	leicht	mittel	schwer																																																								
Vordergrundsegmentierung			X																																																								
Bewegungsanalyse	X																																																										
Personendetektion	X																																																										
Tracker			X																																																								
Gesichtsdetektion & -tracking	X																																																										
Kopf-/Kopfschulterdetektion	X																																																										
	<table border="1"> <thead> <tr> <th>Verfahren</th> <th>leicht</th> <th>mittel</th> <th>schwer</th> </tr> </thead> <tbody> <tr><td>Vordergrundsegmentierung</td><td>X</td><td></td><td></td></tr> <tr><td>Bewegungsanalyse</td><td>X</td><td></td><td></td></tr> <tr><td>Personendetektion</td><td></td><td>X</td><td></td></tr> <tr><td>Tracker</td><td></td><td>X</td><td></td></tr> <tr><td>Gesichtsdetektion & -tracking</td><td>X</td><td></td><td></td></tr> <tr><td>Kopf-/Kopfschulterdetektion</td><td>X</td><td></td><td></td></tr> </tbody> </table>	Verfahren	leicht	mittel	schwer	Vordergrundsegmentierung	X			Bewegungsanalyse	X			Personendetektion		X		Tracker		X		Gesichtsdetektion & -tracking	X			Kopf-/Kopfschulterdetektion	X				<table border="1"> <thead> <tr> <th>Verfahren</th> <th>leicht</th> <th>mittel</th> <th>schwer</th> </tr> </thead> <tbody> <tr><td>Vordergrundsegmentierung</td><td>X</td><td></td><td></td></tr> <tr><td>Bewegungsanalyse</td><td>X</td><td></td><td></td></tr> <tr><td>Personendetektion</td><td>X</td><td></td><td></td></tr> <tr><td>Tracker</td><td>X</td><td></td><td></td></tr> <tr><td>Gesichtsdetektion & -tracking</td><td></td><td></td><td>X</td></tr> <tr><td>Kopf-/Kopfschulterdetektion</td><td></td><td>X</td><td></td></tr> </tbody> </table>	Verfahren	leicht	mittel	schwer	Vordergrundsegmentierung	X			Bewegungsanalyse	X			Personendetektion	X			Tracker	X			Gesichtsdetektion & -tracking			X	Kopf-/Kopfschulterdetektion		X	
Verfahren	leicht	mittel	schwer																																																								
Vordergrundsegmentierung	X																																																										
Bewegungsanalyse	X																																																										
Personendetektion		X																																																									
Tracker		X																																																									
Gesichtsdetektion & -tracking	X																																																										
Kopf-/Kopfschulterdetektion	X																																																										
Verfahren	leicht	mittel	schwer																																																								
Vordergrundsegmentierung	X																																																										
Bewegungsanalyse	X																																																										
Personendetektion	X																																																										
Tracker	X																																																										
Gesichtsdetektion & -tracking			X																																																								
Kopf-/Kopfschulterdetektion		X																																																									
	<table border="1"> <thead> <tr> <th>Verfahren</th> <th>leicht</th> <th>mittel</th> <th>schwer</th> </tr> </thead> <tbody> <tr><td>Vordergrundsegmentierung</td><td></td><td></td><td>X</td></tr> <tr><td>Bewegungsanalyse</td><td></td><td></td><td>X</td></tr> <tr><td>Personendetektion</td><td>(X)</td><td></td><td></td></tr> <tr><td>Tracker</td><td>(X)</td><td></td><td></td></tr> <tr><td>Gesichtsdetektion & -tracking</td><td>(X)</td><td></td><td></td></tr> <tr><td>Kopf-/Kopfschulterdetektion</td><td>(X)</td><td></td><td></td></tr> </tbody> </table>	Verfahren	leicht	mittel	schwer	Vordergrundsegmentierung			X	Bewegungsanalyse			X	Personendetektion	(X)			Tracker	(X)			Gesichtsdetektion & -tracking	(X)			Kopf-/Kopfschulterdetektion	(X)				<table border="1"> <thead> <tr> <th>Verfahren</th> <th>leicht</th> <th>mittel</th> <th>schwer</th> </tr> </thead> <tbody> <tr><td>Vordergrundsegmentierung</td><td></td><td></td><td>X</td></tr> <tr><td>Bewegungsanalyse</td><td></td><td>X</td><td></td></tr> <tr><td>Personendetektion</td><td></td><td>X</td><td></td></tr> <tr><td>Tracker</td><td></td><td>X</td><td></td></tr> <tr><td>Gesichtsdetektion & -tracking</td><td></td><td>X</td><td></td></tr> <tr><td>Kopf-/Kopfschulterdetektion</td><td></td><td>X</td><td></td></tr> </tbody> </table>	Verfahren	leicht	mittel	schwer	Vordergrundsegmentierung			X	Bewegungsanalyse		X		Personendetektion		X		Tracker		X		Gesichtsdetektion & -tracking		X		Kopf-/Kopfschulterdetektion		X	
Verfahren	leicht	mittel	schwer																																																								
Vordergrundsegmentierung			X																																																								
Bewegungsanalyse			X																																																								
Personendetektion	(X)																																																										
Tracker	(X)																																																										
Gesichtsdetektion & -tracking	(X)																																																										
Kopf-/Kopfschulterdetektion	(X)																																																										
Verfahren	leicht	mittel	schwer																																																								
Vordergrundsegmentierung			X																																																								
Bewegungsanalyse		X																																																									
Personendetektion		X																																																									
Tracker		X																																																									
Gesichtsdetektion & -tracking		X																																																									
Kopf-/Kopfschulterdetektion		X																																																									
	<table border="1"> <thead> <tr> <th>Verfahren</th> <th>leicht</th> <th>mittel</th> <th>schwer</th> </tr> </thead> <tbody> <tr><td>Vordergrundsegmentierung</td><td></td><td></td><td>X</td></tr> <tr><td>Bewegungsanalyse</td><td></td><td>X</td><td></td></tr> <tr><td>Personendetektion</td><td></td><td>X</td><td></td></tr> <tr><td>Tracker</td><td></td><td>X</td><td></td></tr> <tr><td>Gesichtsdetektion & -tracking</td><td></td><td>X</td><td></td></tr> <tr><td>Kopf-/Kopfschulterdetektion</td><td></td><td>X</td><td></td></tr> </tbody> </table>	Verfahren	leicht	mittel	schwer	Vordergrundsegmentierung			X	Bewegungsanalyse		X		Personendetektion		X		Tracker		X		Gesichtsdetektion & -tracking		X		Kopf-/Kopfschulterdetektion		X																															
Verfahren	leicht	mittel	schwer																																																								
Vordergrundsegmentierung			X																																																								
Bewegungsanalyse		X																																																									
Personendetektion		X																																																									
Tracker		X																																																									
Gesichtsdetektion & -tracking		X																																																									
Kopf-/Kopfschulterdetektion		X																																																									

Abbildung 28: Übersicht der ausgewählten Sequenzen inklusive Kategorisierung der einzelnen Verfahren.

Die Ergebnisse der Analyse sind in den folgenden Grafiken dargestellt. Diese erlauben objektive Vergleiche der einzelnen Verfahren. Dies soll anhand von folgendem Beispiel verdeutlicht werden (bezogen auf Abbildung 29): Der GPU-HOG der HRW weist hier im Vergleich zum Ground-HOG der TU-I viele Falsch-Positiv-Detektionen auf (GPU-HOG: 472; Ground-HOG: 50). Die Anzahl an Falsch-Negativen beträgt allerdings beim GPU-HOG lediglich 6, beim Ground-HOG dagegen 1199. Insgesamt führt dies zu einem hohen MODA-Wert beim GPU-HOG (0,94) und einem niedrigen MODA-Wert beim Ground-HOG (0,43). Der GPU-HOG schneidet also besser ab, da dem Ground-HOG in dieser Sequenz sehr viele Personen verloren gehen.

EASC C2 - Detektor

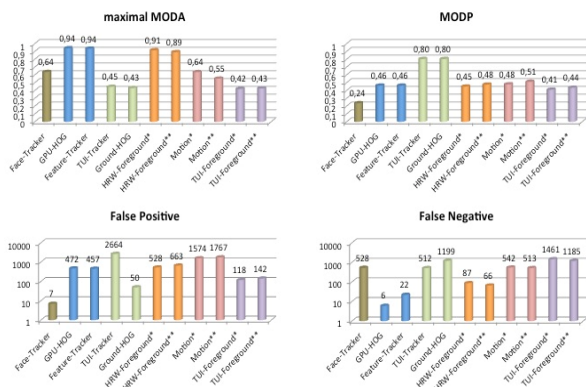


Abbildung 29: Evaluationsergebnisse der Detektionsverfahren auf der Sequenz EASC C2. (* Alle Körperteile wurden berücksichtigt, ** Nur Ganzkörperlabel).

EASC C2 - Tracker

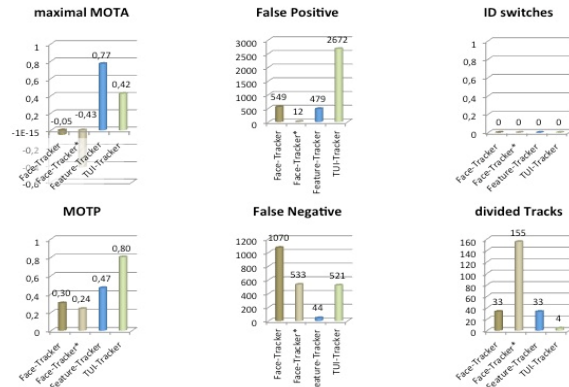


Abbildung 30: Evaluationsergebnisse der Trackingverfahren auf der Sequenz EASC C2. (Ziel-MOTP: min. 0,0).

EASC C7/Erfurt C2 - Detektor

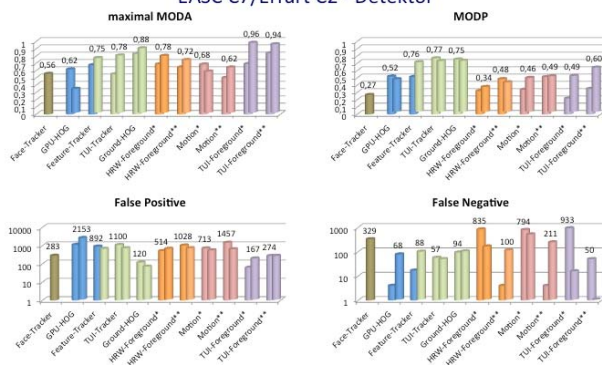


Abbildung 31: Evaluationsergebnisse der Detektionsverfahren auf den Sequenzen EASC C7 und Erfurt C2 (* Alle Körperteile wurden berücksichtigt, ** Nur Ganzkörperlabel).

EASC C7/Erfurt C2 - Tracker

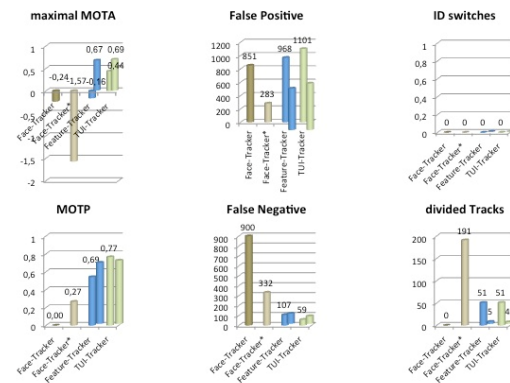


Abbildung 32: Evaluationsergebnisse der Trackingverfahren auf den Sequenzen EASC C7 und C2. (* Ziel-MOTP: min. 0,0).

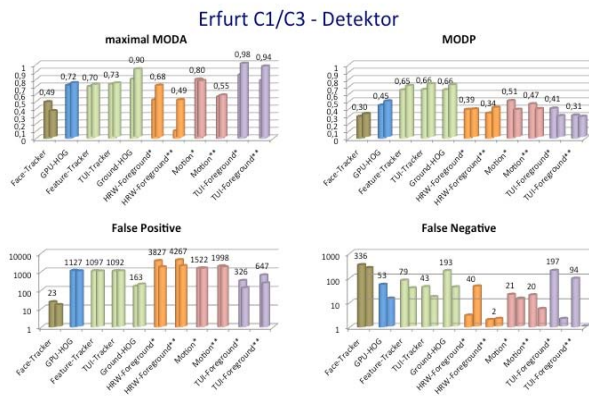


Abbildung 34: Evaluationsergebnisse der Detektionsverfahren auf den Sequenzen Erfurt C1 und C3. (* Alle Körperteile wurden berücksichtigt, ** Nur Ganzkörperlabel).

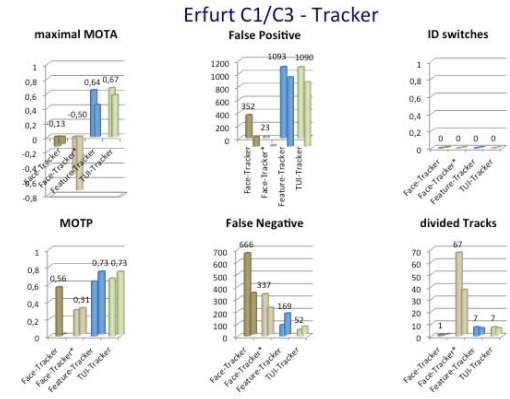


Abbildung 33: Evaluationsergebnisse der Trackingverfahren auf den Sequenzen Erfurt C1 und C3. (* Ziel-MOTP: min. 0,0).

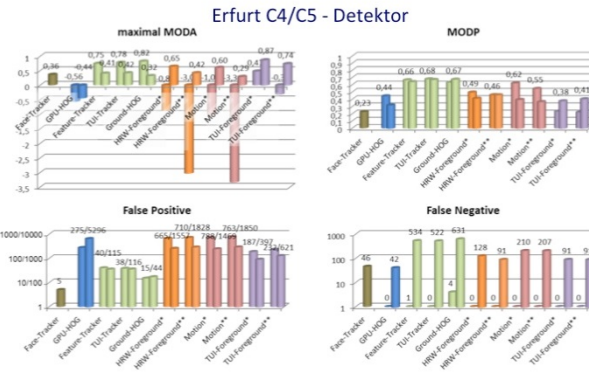


Abbildung 35: Evaluationsergebnisse der Detektionsverfahren auf den Sequenzen Erfurt C4 und C5. (* Alle Körperteile wurden berücksichtigt, ** Nur Ganzkörperlabel).

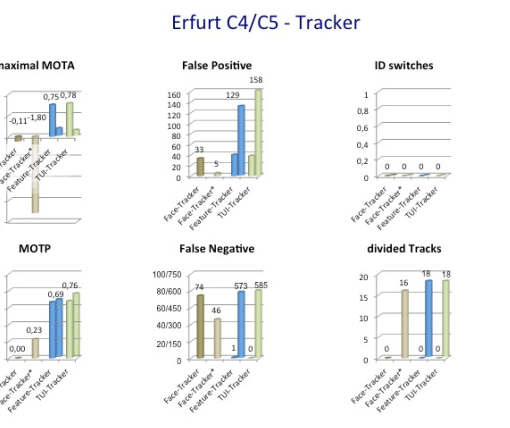


Abbildung 36: Evaluationsergebnisse der Trackingverfahren auf den Sequenzen Erfurt C4 und C5. (* Ziel-MOTP: min. 0,0).

Eine Analyse der verschiedenen Diagramme ergibt, dass die einzelnen Verfahren auf verschiedenen Sequenzen unterschiedlich gute Ergebnisse liefern und somit der Fusionsansatz trägt (siehe Kapitel 5.2.9). Durch die Kombination der verschiedenen Verfahren ist es möglich, kameraspezifisch die Verfahren so zu fusionieren, dass die bestmöglichen Resultate erzielt werden können. Die Detektionsgüte und Robustheit des Gesamtsystems wird so erheblich erhöht.

5.3.4 SWIR/NIR-KAMERAS

Im Rahmen der Evaluation der Wiedererkennungsverfahren wurde die These bestärkt, dass eine Wiedererkennung bei Personen mit ähnlicher Kleidung oft sehr schwierig ist. Eine eindeutige Zuordnung ist nur durch das Gesicht möglich. Dieses ist aber oft nicht oder nur in unzureichender Auflösung im Videomaterial vorhanden, so dass der Suchraum bei vielen ähnlich gekleideten Personen nur unzureichend durch die auf der Kleidung operierenden Verfahren eingeschränkt werden kann. Um dieses Problem zu reduzieren wurden weitere, diskriminierende Merkmale untersucht. Da infrarote Spektren häufig im Überwachungsbereich eingesetzt werden, wurden zunächst IR-Kameras untersucht. Diese Kameras visualisieren unter anderem das Wärmebild einer Person, welches sich durch die unterschiedlichen Isolationseigenschaften der Kleidung und der individuellen Körperwärme unterscheiden. Allerdings variiert dieses Spektrum auch stark durch die

Umgebungstemperatur. Eine Person, welche sich von einem warmen in einen kalten Bereich oder andersherum bewegt, verändert damit auch ihre Wärmebildsignatur erheblich (z. B. Wechsel Außen- und Innenbereiche). Ein weiteres Spektrum, welches zwischen dem typischen IR-Spektrum und dem Spektrum des sichtbaren Lichtes liegt ist das nahinfrarote Spektrum (NIR oder auch SWIR) (Abbildung 37, Links).

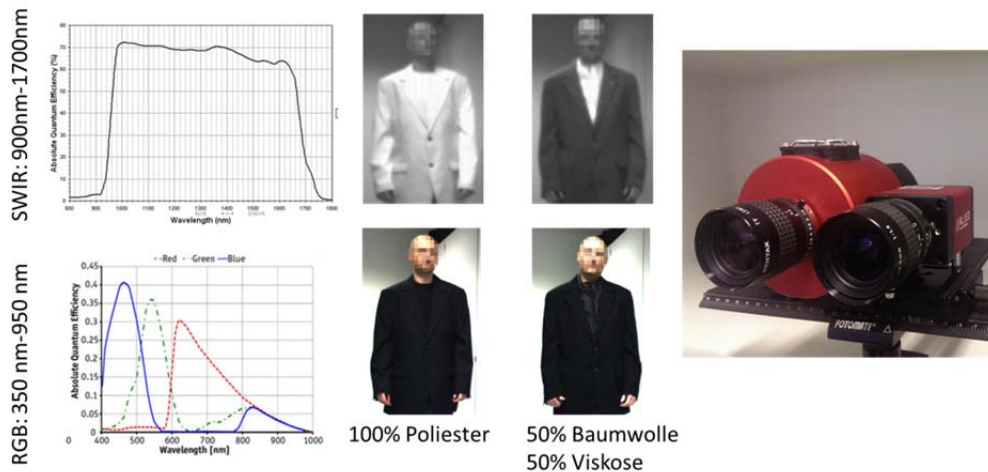


Abbildung 37: NIR/SWIR-Kameras: Links: Wellenspektrum einer RGB- (unten) und einer SWIR-Kamera (oben); Mitte oben: Bild zweier schwarzer Anzüge als NIR/SWIR-Aufnahme; Mitte unten: dieselben schwarzen Anzüge wie darüber als RGB-Aufnahme; Rechts: Kombination SWIR/NIR-Kamera und RGB-Kamera

In diesem Wellenspektrum werden Materialeigenschaften sichtbar, da verschiedene Materialien dieses Spektrum, ähnlich wie beim sichtbaren Licht, unterschiedlich stark reflektieren [21]. Einige Materialien reflektieren dieses Lichtspektrum nahezu vollständig, andere absorbieren die Strahlung ganz oder lassen diese ungefiltert durch. Da diese Eigenschaft von dem Material und nicht der Farbe abhängig ist, bietet dieses Spektrum einen zusätzlichen Merkmalsraum um Personen zu unterscheiden. Kleidungsstücke, welche im Spektrum des sichtbaren Lichtes gleich erscheinen, unterscheiden sich im NIR/SWIR-Spektrum teilweise stark. Das beschriebene Spektrum liefert weitere zusätzliche individuelle Merkmale (Abbildung 37 Mitte).

Auf Grund der mittels NIR/SWIR-Kameras ermöglichten Erweiterung des Merkmalsraums wurden diese weiter untersucht, wozu eine entsprechende Kamera angeschafft wurde. Bei der Analyse des erweiterten Merkmalsraums stand vor allem die Übertragbarkeit der bisherigen Verfahren auf diesen Sensortyp im Fokus. Da die Helligkeit und der Kontrast einer NIR/SWIR-Aufnahme von der Lichtmenge abhängt, welche auf die aufgezeichnete Szene fällt, ist auch bei diesen Aufnahmen eine Beleuchtungskorrektur notwendig. Dabei lässt sich das entwickelte Verfahren der nichtlinearen Abbildung der kamerainternen Pixelrepräsentation (12 Bit) auf ein typisches 8 Bit-Bild direkt übertragen. Auch die Detektion von Personen ist hier direkt anwendbar. Da eine NIR/SWIR-Aufnahme auf Grund des differierenden Spektralbereichs andere Reflexionseigenschaften wie im sichtbaren Wellenspektrum aufweist, erscheinen die aufgenommenen Bilder immer leicht verwaschen. Dadurch verringert sich die Detektionsgüte der auf Kantenbildern arbeitenden Verfahren. Eine Erweiterung der Datenbank für das Training des im Detektor verwendeten Klassifikators um Beispielbilder aus dem NIR-Spektralbereich kann führt zu einer Verbesserung, dennoch fehlen die im sichtbaren Licht so stark ausgeprägten Kanten. Bei einer parallelen Anordnung einer RGB- und einer SWIR/NIR-Kamera (Abbildung 37 Rechts) lässt sich das im RGB-Bild ermittelte

Detektionsfenster aber leicht auf das NIR/SWIR-Bild abbilden. Weitere Merkmale basieren auf der Farbe und Textur eines Bildes. Da eine NIR/SWIR-Aufnahme aber nur ein Grauwertbild darstellt, sind diese nicht direkt übertragbar. Allerdings lassen sich die Ansätze übertragen, was dazu führt, dass in dem NIR/SWIR-Bild der mittlere Grauwert und ein 16bin-Histogramm über die Grauwerte ermittelt wird. Die horizontalen und vertikalen Texturanteile lassen sich dagegen äquivalent zu einem RGB-Bild ermitteln. Die individuellen Merkmale lassen sich ebenfalls auf eine SWIR/NIR-Aufnahme übertragen. Die einzige Ausnahme bildet dabei die Berechnung der Aufmerksamkeitskarte auf Basis des Farbwertes (Abbildung 38).

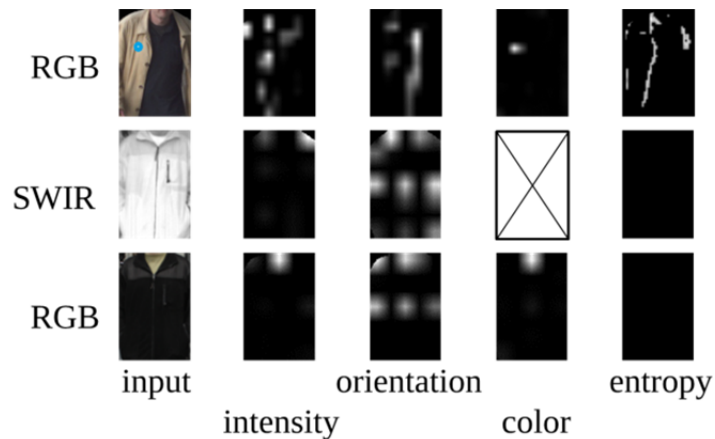


Abbildung 38: Aufmerksamkeitskarten RGB vs. NIR/SWIR: Die Aufmerksamkeitskarten zur Ermittlung der individuellen Merkmale lassen sich bis auf die farbbasierte Karte auch für NIR/SWIR-Bilder bestimmen.

6 ABSCHLUSSDEMONSTRATOR

Ein übergeordnetes Ziel im Rahmen des APFeI-Projektes war die Realisierung eines finalen Demonstrators, um die Leistungsfähigkeit der erarbeiteten Verfahren und die erfolgreiche Umsetzung der geplanten Ziele demonstrieren zu können. Hierzu wurde seitens des Projektpartners TU Ilmenau [29] eine graphische Benutzeroberfläche entwickelt, die die Ergebnisse der realisierten Verfahren der technischen Partner HRW, L-1 Identity Solutions AG und TU Ilmenau visualisiert und diese integriert. Weitere Teile des Demonstrators bilden die von L-1 Identity Solutions AG entwickelte Datenbank und der von der HRW entwickelte Videosever. Jede Kamera ist dabei mit einem Videosever verknüpft, der auf Anfrage die Bilder der Kameras liefert. Die entwickelten Verfahren der einzelnen Partner verarbeiten diese Bilder und schreiben die Ergebnisse in die zentrale Datenbank (Liveanalyse, dezentral). Verfahren zur Anforderungsanalyse (Wiedererkennung von Personen) greifen über das Netzwerk hierauf zu und schreiben ihre Ergebnisse wiederum in die zentrale Datenbank (Rückwärtsanalyse, zentral). Die graphische Benutzeroberfläche stellt dem Operator verschiedene Funktionen bereit, die eine Szenenanalyse basierend auf den einzelnen Verfahren der technischen Partner erlauben. Abbildung 40 zeigt die graphische Benutzeroberfläche. Eine Übersicht über die Kommunikationsstruktur wird in Abbildung 39 verdeutlicht.

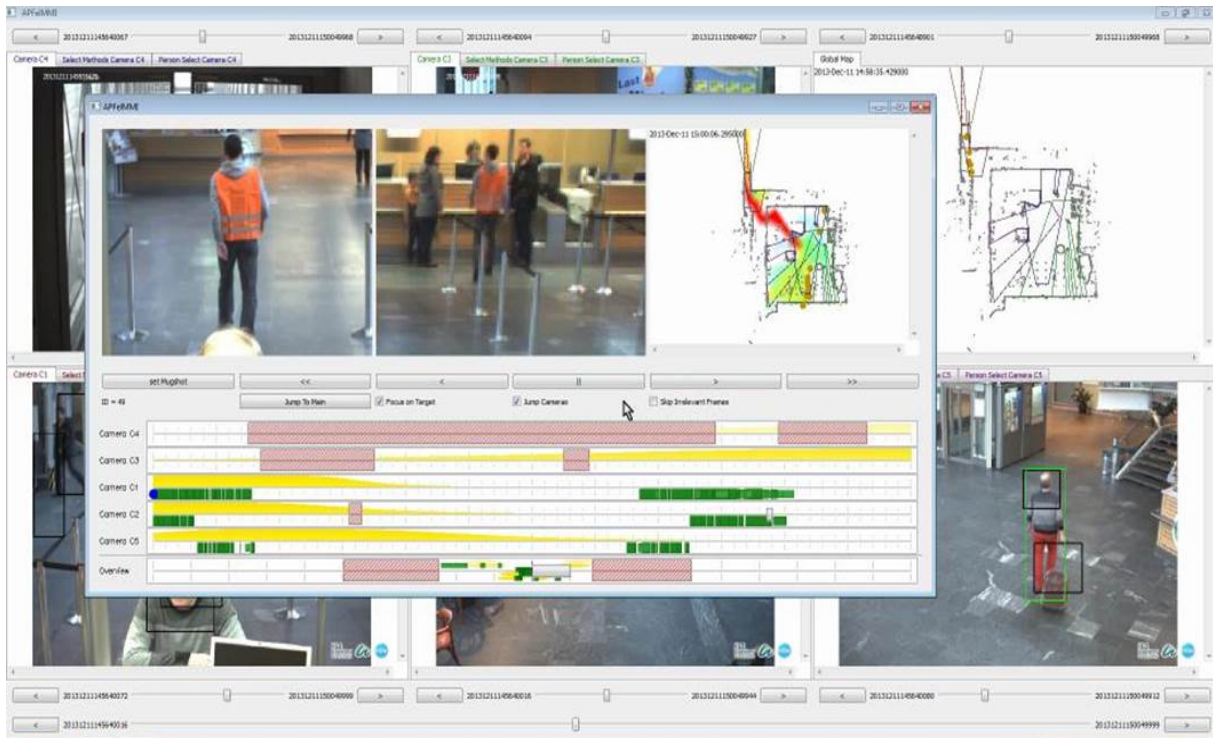


Abbildung 40: Abbildung der graphischen Benutzeroberfläche. Dargestellt werden die einzelnen Kameras, eine Gebäudeübersicht sowie ein Analysefenster (Vordergrund).

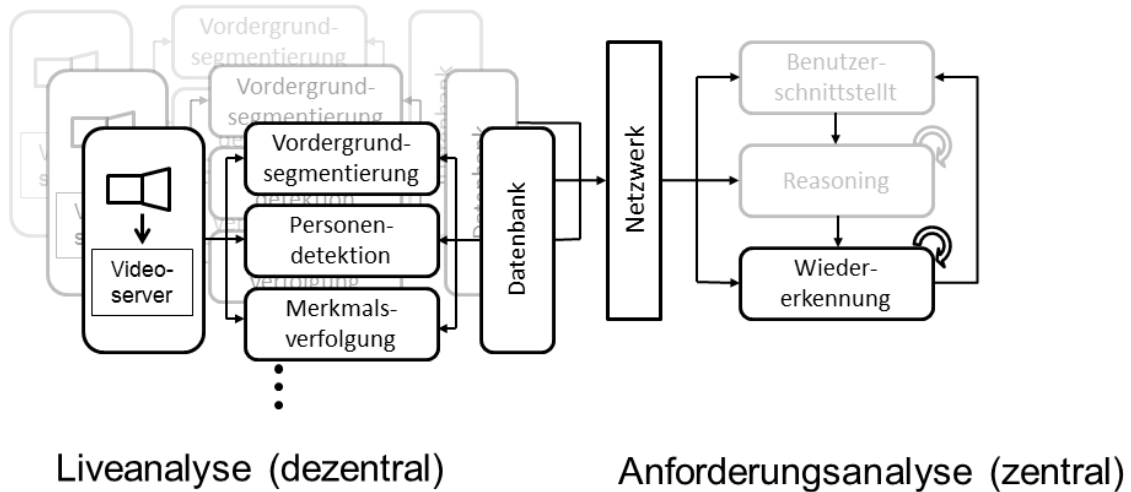


Abbildung 39: Übersicht der Kommunikationsstruktur.

7 LITERATURVERZEICHNIS

- [1] U. Handmann, S. Hommel, M. Brauckmann und M. Dose, „Face detection and person identification on mobile platforms,“ in *Towards Service Robots for Everyday Environments - Springer Tracts in Advanced Robotics (STAR)*, Berlin, Heidelberg, Germany, Springer Verlag, 2012, pp. 227-234.

- [2] G. von Wichert, U. Handmann, C. Klimowicz, W. Neubauer, T. Wösch, G. Lawitzky, R. Caspari, H. J. Heger, P. Witschel und M. Rinne, „The Robotic Bar – An Integrated Demonstration of a Robotic Assistant,“ in *Advances in Human-Robot Interaction*, Berlin Heidelberg, Springer, 2005, pp. 359-370.
- [3] C. Castillo und C. Chang, „An Approach to Vision-Based Person Detection in Robotic Applications,“ in *Pattern Recognition and Image Analysis*, Berlin Heidelberg, Springer, 2005, pp. 209-216.
- [4] M. Lube, G. D. Tipaldi und K. O. Arras, „Spatially grounded multi-hypothesis tracking of people,“ in *In Proceedings of the IEEE ICRA 2009 Workshop on People Detection and Tracking*, Kobe, Japan, 2009.
- [5] D. M. Gavrila, „Pedestrian Detection from a Moving Vehicle,“ in *Computer Vision - ECCV 2000*, Berlin Heidelberg, Springer, 2000, pp. 37-49.
- [6] C. Stauffer und W. E. L. Grimson, „Adaptive background mixture models for real-time tracking,“ in *Computer Vision and Pattern Recognition*, IEEE, 1999, pp. 246-252.
- [7] C. P. Papageorgiou, M. Oren und T. Poggio, „A general framework for object detection,“ in *Sixth International Conference on Computer Vision*, IEEE, 1998, pp. 555-562.
- [8] D. M. Gavrila, „Multi-feature hierarchical template matching using distance transforms,“ in *Fourteenth International Conference on Pattern Recognition*, IEEE, 1998, pp. 439-444.
- [9] D. Crandall, P. Felzenszwalb und D. Huttenlocher, „Spatial priors for part-based recognition using statistical models,“ in *Computer Vision and Pattern Recognition*, IEEE, 2005, pp. 10-17.
- [10] B. Leibe, A. Leonardis und B. Schiele, „Combined Object Categorization and Segmentation With An Implicit Shape Model,“ in *In ECCV workshop on statistical learning in computer vision*, IEEE, 2004, pp. 17-32.
- [11] P. Viola und M. Jones, „Robust Real-time Object Detection,“ in *International Journal of Computer Vision*, 2001.
- [12] P. Viola und M. Jones, „Rapid object detection using a boosted cascade of simple features,“ in *Computer Vision and Pattern Recognition*, IEEE, 2001, pp. 511-518.
- [13] P. Viola, M. Jones und D. Snow, „Detecting pedestrians using patterns of motion and appearance,“ in *Ninth IEEE International Conference on Computer Vision*, IEEE, 2003, pp. 734-741.
- [14] P. Viola und M. Jones, „Image Analysis via the General Theory of Moments,“ in *International Journal of Computer Vision*, 2004, pp. 137-154.
- [15] N. Dalal und B. Triggs, „Histograms of oriented gradients for human detection,“ in *Computer Vision and Pattern Recognition*, IEEE, 2005, pp. 886-893.
- [16] K. Mikolajczyk, C. Schmid und A. Ziss, „Human Detection Based on a Probabilistic Assembly

- of Robust Part Detectors," in *Computer Vision - ECCV 2004*, Springer, 2004, pp. 128-142.
- [17] B. Wu und R. Nevatia, „Improving Part based Object Detection by Unsupervised, Online Boosting," in *Computer Vision and Pattern Recognition*, IEEE, 2007, pp. 1-8.
- [18] B. Wo und R. Nevatia, „Simultaneous object detection and segmentation by boosting," in *Computer Vision and Pattern Recognition*, IEEE, 2007, pp. 1-8.
- [19] S. Hommel, M. Grimm, U. Handmann, V. Voges und U. Weigmann, „An Intelligent System Architecture for Multi-Camera Human Tracking at Airports," in *13th IEEE International Symposium on Computational Intelligence and Informatics*, Budapest, 2012.
- [20] M. Grimm, S. Hommel, U. Handmann, V. Voges und U. Weigmann, „Multi-Camera Human Tracking At Airports Based on an Intelligent System Architecture," in *7th Future Security Conference*, Bonn, 2012.
- [21] S. Hommel, D. Malysiak und U. Handmann, „Model of Human Clothes based on Saliency Maps," in *14th IEEE International Symposium on Computational Intelligence and Informatics*, Budapest, 2013.
- [22] M. Grimm, S. Hommel, U. Handmann, V. Voges und U. Weigmann, „Intelligent Support of Video Surveillance At Airports," in *7th Future Security Conference*, Bonn, 2012.
- [23] S. Hommel, D. Malysiak und U. Handmann, „Efficient people re-identification based on models of human clothes," in *15th IEEE International Symposium on Computational Intelligence and Informatics*, Budapest, Hungary, 2014.
- [24] D. Malysiak und U. Handmann, „An efficient framework for distributed computing in heterogeneous beowulf clusters and cluster-management," in *15th IEEE International Symposium on Computational Intelligence and Informatics*, Budapest, Hungary, 2014.
- [25] D. Malysiak und U. Handmann, „An algorithmic skeleton for massively parallelized mean shift computation with applications to GPU architectures," in *15th IEEE International Symposium on Computational Intelligence and Informatics*, Budapest, Hungary, 2014.
- [26] L. Itti, C. Koch und E. Niebur, „A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE transactions on pattern analysis and machine intelligence*, pp. 1254-1259, November 1998.
- [27] U. Handmann, T. Kalinke, C. Tzomakas, M. Werner und W. von Seelen, „An Image Processing System for Driver Assistance," *Image and Vision Computing, Elsevier Science*, Bd. 18, Nr. 5, pp. 367 - 376, 2000.
- [28] T. Kalinke und W. von Seelen, „Entropie als Maß des lokalen Informationsgehalts in Bildern zur Realisierung einer Aufmerksamkeitssteuerung," *DAGM*, 1996.

Impressum

Technical Report 14-01

ISSN: 2199-9937

1. Auflage, 30.09.2014

© Institut Informatik, Hochschule Ruhr West, Germany

Anschrift

Institut Informatik

Hochschule Ruhr West

Lützowstraße 5

46236 Bottrop