Freie Universität Berlin

Fachbereich Mathematik und Informatik
Institut für Informatik
Softwareprojekt Mobilkommunikation

Vordergrundextraktion auf Basis von Hidden Markov Modellen

Marcin Nawrocki
Patrick Mertes
Hinnerk van Bruinehsen

Betreuer: Alexandra Danilkina

28. Februar 2014

Inhaltsverzeichnis

1 Einleitung

Dieses Kapitel gibt eine Einführung in das Projekt SAFEST[1] und erläutert die vorliegende Problemstellung, welche im Rahmen des Softwareprojektes Mobilkommunikation im Wintersemester 13/14 bearbeitet wurde.

1.1 Kontext und Motivation

SAFEST[1] ist ein deutsch-französisches Projekt mit der Zielsetzung, die Sicherheit an öffentlichen Plätzen und kritischen Infrastrukturen durch Realisierung eines Sensornetzwerkes mit Infrarotkameras zu erhöhen. Um dieses Ziel zu erreichen, überwacht das Sensornetzwerk ein vorgegebenes Areal und verarbeitet die gewonnenen Mittelinfrarotbilder mit Hilfe von Algorithmen in Hinblick auf verschiedene Parameter. Der eigentliche Gewinn an Sicherheit soll dadurch erfolgen, dass automatisiert auf Grundlage der ermittelten Parameter die Dichte der sich in dem Areal aufhaltenden Personen ermittelt wird, woraus wiederum ein Rückschluss auf die Wahrscheinlichkeit des Auftretens einer Massenpanik gezogen werden soll.
Der erste wichtige Schritt bei dieser Analyse ist die automatisiere Erkennung und Zählung von sich im Bildbereich befindlichen Personen. Um diesen Schritt zu ermöglichen, muss zunächst der Hintergrund vom Vordergrund des Bildes getrennt werden. Die verwendete Kamera stellt wärmere Bereiche heller und kältere Bereiche dunkler dar. Daraus folgt, dass Personen erwartungsgemäß eher weiß und Hintergrund erwartungsgemäß eher schwarz dargestellt werden. Allerdings kalibiert sich die Kamera regelmäßig neu, um den gesamten Graustufenverlauf zur darstellung der Wärme zu nutzen. Daraus folgt, dass mitunter auch sehr kalte Bereiche sehr hell dargestellt werden, wenn diese die wärmsten von der Kamera erfassten Temperaturen besitzen. Aus dieser Eigenschaft der Kamera folgt somit auch, dass ein naiver Ansatz wie eine feste Zuordnung von Temperatur und Graustufenwert nicht möglich ist. Ein weiteres Hindernis bei der einfachen Zuordnung von Temperaturwerten zu Vorder- oder Hintergrund ist die Tatsache, dass Kleidung (besonders zum Beispiel dickere Jacken) die Temperatur sehr gut nach außen isolieren und damit Teile von Personen sehr dunkel auf dem Kamerabild erscheinen lassen.
Die Trennung von Vorder- und Hintergrund kann daher nicht allein aufgrund der dargestellten Temperatur erfolgen, sondern muss als zweiten Faktor die Bewegung mit einbeziehen. Personen sind also in der Darstellung der Regel nach helle, bewegte Objekte.
Nach erfolgter Entfernung des Hintergrunds folgt als nächster Schritt in der Analyse die automatisierte Bestimmung der auf dem Bild sich befindlichen Personen. Wird diese Analyse erfolgreich ausgeführt, so muss nur noch dieser Wert an eine Kontrolleinheit übertragen werden, nicht aber das eigentliche Bild, was im Hinblick auf auf den Datenschutz sehr wichtig ist.
Eigentlich für Objekterkennung bewährte Algorithmen wie Histogramm orientierte Gradienten (HOG)[2] und die Mischung gausscher Verteilungsdichten (MOG)[3], können sich auf die Eigenschaften von Vorder- und Hintergrund, insbesondere die durch die Kamerakalibrierung entstehende Dynamik, nicht so einstellen, dass sie befriedigende Ergebnisse liefern.

Daher soll ein neuer Algorithmus entwickelt und evaluiert werden, der über die Fähigkeit verfügt, sich der Dynamik anzupassen, um das Problem zu lösen.

1.2 Aufgabenbeschreibung

Das Ziel unseres Softwareprojektes ist es, ein Verfahren zu entwickeln, welches das Bild einer Infrarotkamera in Vorder- und Hintergrund trennt. Die genauen Anforderungen für diesen Prozess werden in Kapitel 2.1 beschrieben. Hierbei soll die allgemeine Verwendbarkeit von Hidden-Markov-Modellen(HMM)[4] und der Diskreten Kosinustransformation(DCT)[5] gezeigt werden. Die theoretischen Grundlagen werden hierzu im Kapitel 2 ff. erläutert. Schlussendlich soll eine vergleichende Evaluation durchgeführt werden, in dem das entwickelte Verfahren gegen ein bereits vorliegendes Histrogramm-basiertes Verfahren antreten muss(vgl. Kapitel 4).
Da bereits ein Algorithmus zum Zählen von Personen vorliegt, muss dieser nicht innerhalb dieses Projektes entworfen beziehungsweise implementiert werden.

2 Entwurf eines Vordergrundextraktors

In diesem Kapitel wird der nötige, theoretische Hintergrund herausgearbeitet und unser Entwurf gemäß den Anforderungen vorgestellt. Es wird das Konzept der HMM und der DCT erläutert und deren Verknüpfung eingeführt.

2.1 Anforderungsanalyse

Die funktionalen Anforderungen an einen Vordergrundextraktor sind bereits in der Aufgabenbeschreibung (Kapitel 1.2) ausgeführt: Es soll demonstriert werden, dass HMM zur Trennung von Vordergrund und Hintergrund bei stark heterogenen Hintergründen geeignet sind, wie in Abbildung 1 dargestellt.

Abbildung 1: Kamerabild mit heterogenen Hintergund, Vordergrund ist markiert.

Die nichtfunktionalen Anforderungen umfassen vor allem eine einfache Benutzbarkeit durch Automatisierung. Der Ressourcenbedarf soll möglichst gering sein - ein Video wird für gewöhnlich mit 25 Frames pro Sekunde aufgezeichnet, im Idealfall ist eine optimale Analyse jedes einzelnen Frames in Echtzeit möglich. Allerdings ist zu berücksichtigen, dass eine möglichst hohe Korrektheit wichtiger ist, als eine Analyse aller Frames, da relevante Informationen wie Eintritt bzw. Austritt von Personen aufgrund der relativ langsamen Bewegung der Menschen auch von den nachfolgenden Frames erfasst werden. Falls nötig können demnach Frames von der Analyse ausgeschlossen werden. Die Analyse sollte jedoch unabhängig von den Parametern des verwendeten Videomaterials wie beispielsweise unterschiedliche Videoformate oder Auflösungen statt finden.

2.2 Hidden-Markov-Modelle

Die Markov Modelle stammen aus der Wahrscheinlichkeitstheorie und entsprechen einem stochastischem Zustandsautomaten, bei dem die Zustandswechsel gemäß einer Wahrscheinlichkeit statt finden und nicht abhängig von der Vergangenheit sind, sondern nur von dem aktuellen Zustand[4]. Bei den HMM können jene Zustände nicht beobachtet werden, sondern nur die Beobachtungen/Ausgaben, welche während dieses Zustandes

auftreten, sie werden Emissionen genannt. Die Zustandsübergangswahrscheinlichkeiten sind in den HMM somit nicht die einzigen Parameter, die Emissionswahrscheinlichkeiten bilden die zweiten Parameter.
Formal definiert ist ein HMM ein 5er-Tupel $\lambda = (S, V, A, B, \Pi)$ mit:

- $S = s_1, ..., s_n$ sei die Menge aller Zustände
- $V = v_1, ..., v_m$ das Alphabet der möglichen Emissionen
- $A \in \mathbb{R}^{n \times n}$ sei eine Übergangsmatrix zwischen den Zuständen, a_{ij} entspricht der Wahrscheinlichkeit des Übergangs von Zustand s_i in Zustand s_j
- $B \in \mathbb{R}^{n \times m}$ sei eine Beobachtungsmatrix, wobei $b_i(v_j)$ die Wahrscheinlihkeit angibt, im Zustand s_i die Beobachtung (Emission) v_j zu sehen
- $\Pi \in R^n$ die Initialverteilung, Π_i sei die Wahrscheinlichkeit, dass s_i der Startzustand ist

Ein HMM heißt zeitinvariant, wenn sich die Übergangs- und Emissionswahrscheinlichkeiten nicht mit der Zeit ändern.
HMM werden zunehmend in der Literatur zur Sprach-, Schrift und Mustererkennung [6][7] verwendet, da sie mit den probabilistischen Übergängen die Prozesse der echten Welt besser widerspiegeln als deterministische Definitionen, zu dem können die Parameter durch Lernalgorithmen automatisiert bestimmt werden, siehe hierzu Kapitel 2.3.

Zum besseren Verständnis der vorangehenden Definition visualisiert Abbildung 2 ein exemplarisches HMM. Es liegt ein HMM mit 3 Zuständen und 4 Emissionen vor, wobei der Übergang aus jedem Zustand zu jedem Zustand und zusätzlich in jedem Zustand jede Emission möglich ist. In diesem Beispiel liegt eine Gleichverteilung vor, Übergänge zwischen Zuständen haben stets die Wahrscheinlichkeit 1/3, die Emissionen haben stets eine Wahrscheinlichkeit von 1/4. Beschriftung der Kanten aus Übersichtsgründen nicht dargestellt.

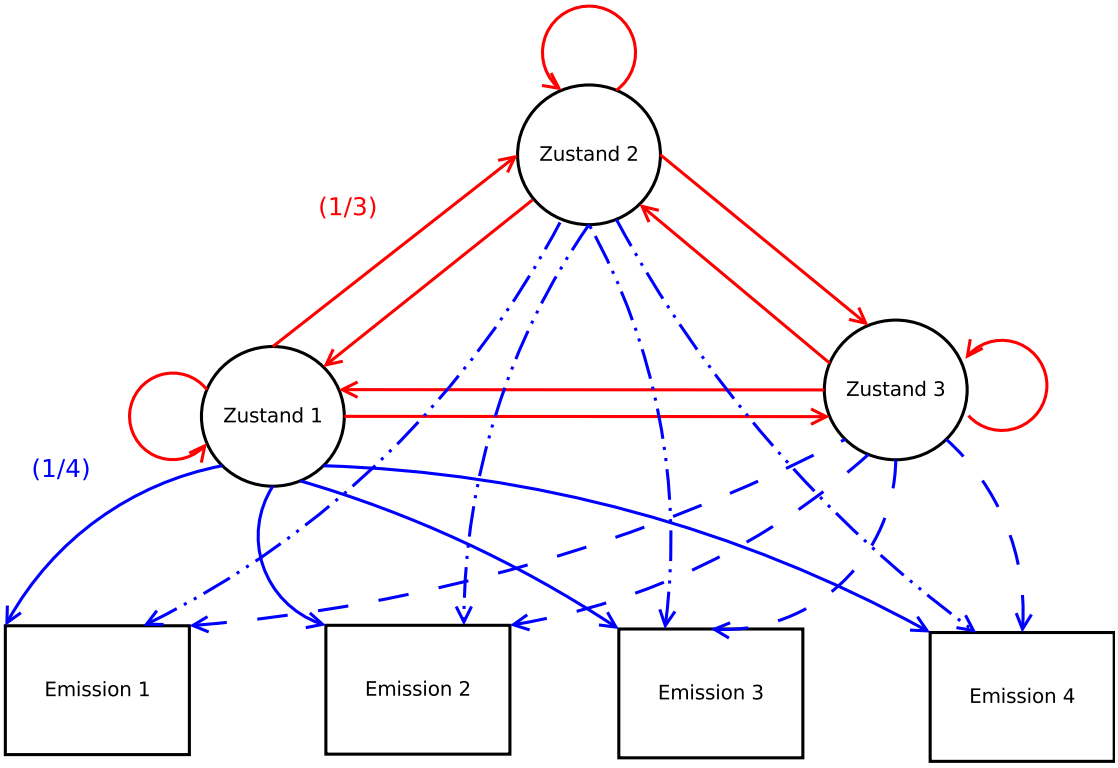

Abbildung 2: Beispiel für ein HMM, welches eine Gleichtverteilung aufweist. Eine Gleichverteilung wird oft initial angenommen und dann ein Lernprozess ausgeführt.

2.3 Gängige Problemstellungen der HMM

Für die HMM wurden bereits viele Algorithmen entwickelt, welche Standardprobleme lösen. Die zwei häufigsten Problemstellungen bei HMMs sind das Lernproblem und das Evaluations-/Decodingproblem.

1. Das Lernproblem bezieht sich auf die Problematik ein HMM anhand von Emissionssequenzen zu trainieren. Gelöst wird dieses Problem durch den Baum-Welch Algorithmus, welcher dazu verwendet wird, unbekannte Parameter, genauer die Übergangs- und Emiossionswahrscheinlichkeiten eines HMM, zu bestimmen. Hierbei handelt es sich um einen erwartungsmaximierenden Algorithmus, welcher anhand von übergebenen Trainingssequenzen die Maximum-Likelihood-Schätzwerte berechnet. Die initialen Werte eines HMMs müssen geschätzt werden, hier wird für gewöhnlich eine Gleichverteilung angenommen, dass heißt das jeder Übergang und jede Emission in einem Zustand gleich wahrscheinlich sind, vergleiche Abbildung 2.

2. Das Evaluations-/Decodingproblem bezieht sich auf die Problemaik aus den ver-

schieden Wahrscheinlichkeiten Aussagen über die Folgezustände oder aber über die Wahrscheinlichkeit bestimmter Zustandsketten zu treffen. Es wird durch den Forward-Algorithmus beziehungsweise den eng verwandten Viterbi-Algorithmus gelöst. Gegeben sei eine Chronik an k-letzten Emissionen, was ist die Wahrscheinlichkeit für eine bestimmte Emission? Und ferner, was ist die wahrscheinlichste Emission? Der Viterbi-Algorithmus berechnet zur Beantwortung dieser Frage die wahrscheinlichste Zustandssequenz, also eine Sequenz die die Wahrscheinlichkeit der übergebenen Emissionssequenz maximiert. Der Forward-Algorithmus ist ein Algorithmus aus der dynamischen Programmierung und optimiert im Gegensatz zum Viterbi nicht rückwirkend die gesamte Zustandssequenz neu, sondern berechnet den aktuell wahrscheinlichsten Zustand auf Basis der zuvor berechneten Zustände und hängt diesen an. Somit ist der Forward Algorithmus zur Laufzeit grundsätzlich schneller, jedoch nicht so genau wie der Viterbi-Algorithmus.

Anhand des gewonnen letzten Zustandes kann nun bestimmt werden, welche Emission als nächstes vorhergesagt wird. Hierbei wird aus den möglichen Emissionen des letzten Zustandes deterministisch die wahrscheinlichste Emission gewählt oder aber gemäß der Wahrscheinlichkeiten gewichtetet gewürfelt. Die zweite Variante ist aufgrund ihrer probabilistischen Natur stärker an den HMM orientiert und sollte daher vorgezogen werden.

2.4 Diskrete Kosinustransformation

Die Diskrete Kosinustransformation (DCT)[5] ist ein Verfahren, welches zur verlustbehafteten Kompression von Daten verwendet wird, wobei die bekannteste Anwendung das Dateiformat JPEG ist. Ähnlich zu der diskreten Fourier Transformation wird eine Information mittels Frequenzen repräsentiert, wobei wie der Name bereits nahe legt nur Kosiunusfunktionen verwendet werden.

Es existieren mehrere Varianten der DCT. Wird jedoch die DCT auf Bildinformationen angewandt, speziell bei der JPEG-Kompression, so wird die zwei-dimensionale Typ II DCT verwendet, diese ist demnach die gängigste Variante. Die DCT erhält eine $n \times n$ Matrix und gibt eine Matrix der selben Größe zurück, wobei im JPEG-Standard 8×8 Pixel Matrizen betrachtet werden. Hierbei wird das Element $[1, 1]$ als DC-Wert bezeichnet und bildet den Durchschnittswert der Farben der betrachteten Matrix; die restlichen 63 Werte der Matrix sind ein Offset zu dem DC-Wert und kodieren somit den Unterschied zum DC-Wert innerhalb der Matrix. Diese Werte werden als AC-Werte bezeichnet.
Formal definiert ist die DCT eine lineare, invertierbare Funktion $f : \mathbb{R}^n \to \mathbb{R}^n$, welche N reellwertige Werte aus $x[n]$ in N reellwertige Werte nach $X[n]$ überführt:

$$X_k = \sum_{n=0 \to N-1} x_n \cos[\tfrac{\Pi}{N} + (n + \tfrac{1}{2})k] \text{ mit } k = 0, ..., N-1$$

Abbildung 3 zeigt das Ergebnis nach der Anwendung der DCT mit Einfärben der gesamten Blöcke in die jeweils berechneten DC-Werte.

Abbildung 3: Links: Originalbild, rechts: Bild nach Anwendung der DCT blockweise nach DC-Wert eingefärbt

2.5 Modellierung eines Eingabealphabets

In unserem Projekt sollen HMM benutzt werden, um eine Entscheidung zu treffen, ob ein Bildbereich zum Vorder- oder zum Hintergrund gehört. Nach dem das generelle Konzept der HMM in Kapitel 2.2 vorgestellt wurde, fehlt jedoch noch die Definition, wie eine Emission aussieht. Das Alphabet der Emissionen wird als Eingabealphabet bezeichnet, da mit dessen Hilfe das System beschrieben werden kann. Hierbei ist zu beachten, dass das Eingabealphabet endlich und diskret sein muss. Zusätzlich sollte das Eingabealphabet aus möglichst wenigen, aussagekräftigen Symbolen bestehen. Die Definition eines Eingabealphabets, also die Zuordnung von Daten zu bestimmten Emissionen, gehört daher in unserem Fall zu der Kernleistung bei der Verwendung der HMM.
Die im Kapitel 2.3 vorgestellte DCT wird von uns verwendet, um ein Eingabealphabet zu erzeugen, wobei wir die DCT auf fest definierte Bereiche anwenden. Wir übernehmen die aus der Bildkompression üblichen 8 x 8 Pixelblöcke zur Einteilung des vorliegenden Bildes, da wir dieses als optimalen Kompromiss zwischen einer zu hohen und niedrigen Granularität ansehen: kleinere Betrachtungen, so zum Beispiel pixelbasierte Verfahren, wären deutlich rechenintensiver und anfälliger auf Bildrauschen und würden daher keine

akkuraten Rückschlüsse bei Veränderungen des Pixelwertes ermöglichen. Bei größeren Betrachtungen wären relevante Veränderungen deutlich schwerer wahrzunehmen, speziell im DC-Wert, da dieser stets den Mittelwert aller Subpixel der Matrix darstellt.

2.5.1 DC-Wert

Da Bildaufnahmen von Mittelinfrarot-Kameras stets schwarzweiß sind (und warme Bereiche heller dargestellt werden als kühle) besitzt ein Pixel genau einen Informationskanal mit einem Grauwert zwischen 0 (schwarz) bis 255 (weiß). Der DC-Wert liegt somit als Mittelwert ebenso in diesem Intervall. Dieser Bereich ist zu groß um als Eingabealphabet für das HMM dienen zu können, da zum Beispiel der Übergang von einem Grauwert von 230 auf 235 nicht aussagekräftig ist, da es sich hierbei um Rauschen oder aber eine Neukalibrierung der Kamera, die sich ja stets auf die gesamte 8×8 Matrix auswirkt, halten kann.

Nimmt man einen kühlen (dunklen) Hintergrund, warmen (hellen) Vordergrund und ein mittelwarmen (grauen) Übergangsbereich an, so kann von einer Dreiteilung des Wärmespektrums ausgegangen werden. Diese Dreiteilung ergibt sich daraus, dass es sehr wahrscheinlich ist, dass sehr helle Bereiche zum Vordergrund und sehr dunkle Bereiche zum Hintergrund gehören. Bei dem aus verschiedenen Grauwerten bestehenden Bereich ist eine einfache Zuordnung allerdings nicht möglich. Eine feinere Einteilung des grauen Clusters ist nicht zielführend, da diese nur Zuordnungen wie *wahrscheinlich Vordergrund* oder *wahrscheinlich Hintergrund* zu ließe und somit keinen bestimmende Information darstellen würde.

Histogramm-basierte Analysen der DC-Werte über mehrere unterschiedliche Videosequenzen über alle Pixelblöcke genormt bestätigen diese Annahme und manifestieren drei Cluster, die wir vereinfacht Cluster schwarz, Cluster grau und Cluster weiß nennen - der Darstellungsfarbe des Wärmespektrums entsprechend. Die Grenzen dieser Cluster sind abhängig von jedem Video und dem betrachteten Umfeld, denn insbesondere die Kälte (und damit die Farbe) des Hintergrundes kann sich bei den jeweiligen Videos stark unterscheiden.
Wir bilden auf Basis dieser Erkenntnis die ersten drei Beobachtungen, wobei, wenn der DC-Wert eines Blockes in das Intervall eines Clusters fällt, wir die entsprechende Beobachtung erstellen:

- DC-Wert im Intervall Cluster schwarz → Beobachtung: DC_BLACK
- DC-Wert im Intervall Cluster grau→ Beobachtung: DC_GREY
- DC-Wert im Intervall Cluster weiß → Beobachtung DC_WHITE

Abbildung 4 zeigt eine exemplarische Einteilung in die drei Cluster. Die X-Achse beschreibt dabei den Grauwert, die Y-Achse die Anzahl Blöcke eines bestimmten Grauwerts. Für die Erstellung des Histogramms wurden sämtliche Grauwerte aller Blöcke eines Testvideos über die gesamte Länge des Videos erfasst.

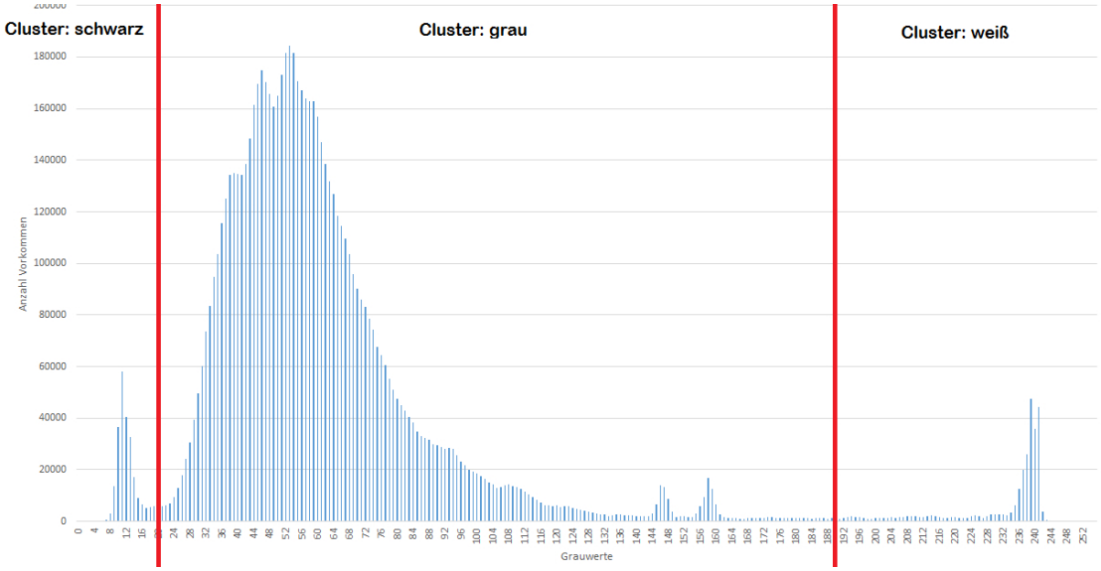

Abbildung 4: Histogramm über alle DC-Werte aller Blöcke einer Videosequenz, Clustergrenzen rot

2.5.2 AC-Wert

Die bei der Anwendung der DCT entstehende 8 x 8 Matrix enthält 63 AC-Werte. Diese bilden demnach den Großteil der in der Matrix kodierten Informationen; die Integration dieser Werte in das Eingabealphabet ist aufgrund der hohen Informationsdichte essenziell.
Eine Histogramm-basierte Betrachtung der AC-Werte analog zu der Betrachtung der DC-Werte ist nicht sinnvoll, da es sich hierbei um Offsets handelt, welche demnach im Histogramm ein Maximum um den Nullwert bilden. Wir erkennen jedoch, das Menschen aufgrund ihrer Wärmeaustrahlung eine kontrastreiche Kante zu dem Hintergrund bilden, läuft demnach ein Mensch durch einen Block, müsste eine deutliche Abweichung einiger AC-Werte zum DC-Wert entstehen. Eine Block-basierte Analyse bestätigt diese Annahme, die Standardabweichung (STD) der AC-Werte bezüglich des DC-Wertes steigt deutlich an, falls eine Person sich im Block befindet beziehungsweise durch diesen hindurch läuft.
Abbildung 5 stellt auf der X-Achse den Zeitverlauf in Frames und auf der Y-Achse den Wert der Standardabweichung für den betrachteten Bock; es lässt sich feststellen, dass die Kurve der Standabweichung stark ausschlägt, falls eine Person durch den Block läuft.

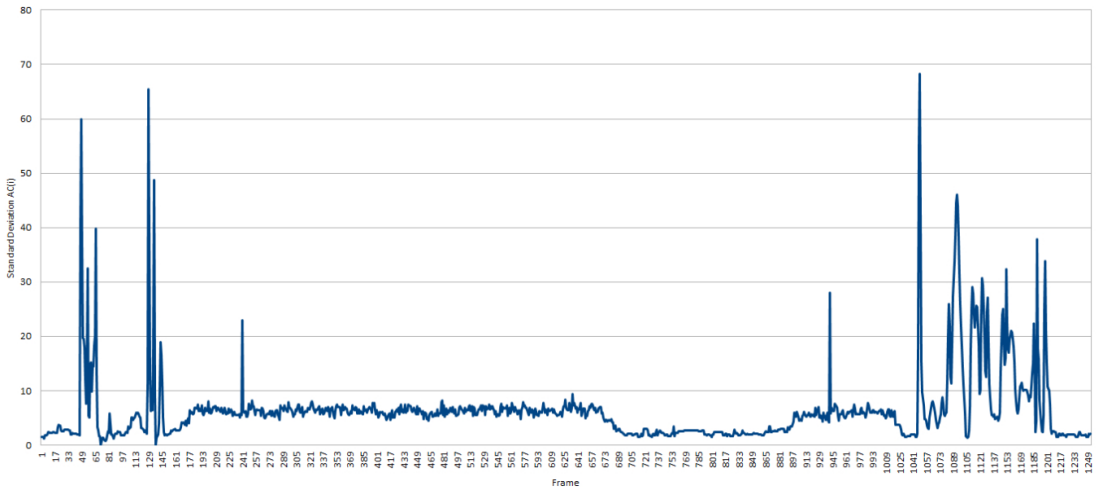

Abbildung 5: Standartabweichung der AC-Werte eines diskreten Blocks über der Zeit

Anhand von empirischen Tests lässt sich ein Schwellwert ermitteln. Falls die Standardabweichung diesen übersteigt, betritt oder verlässt mit hoher Wahrscheinlichkeit eine Person den betrachteten Block. In dem vorgestellten Beispiel liegt dieser bei einer STD von circa 25.
Zusätzlich ist zu beachten, dass die Standardabweichung von wenigen, aber extremen statistischen Ausreißern kaum beeinflusst wird, diese Situation liegt jedoch während der Erkennung von Menschen vor, insbesondere bei Personen, die Kleidung tragen, welche die Wärmeabgabe der Person dämmt und dennoch einige freiliegende Körperteile sichtbar sind. In diesem Fall ist der Großteil der AC-Werte relativ klein und der DC-Wert des betrachteten Blockes für gewöhnlich innerhalb des grauen Intervalls. Um in diesen Situationen den Vordergrund besser zu erkennen, prüfen wir mithilfe eines selbstdefinierten Verfahrens outliers() ob stark-positive (warme) Ausreißer vorliegen, welches den Wahrheitswert true zurück liefert, falls eine gewissen Anzahl von Ausreißern erkannt wird.
Vor allem bei Kleidung tragenden Personen führt dies dazu, dass nicht nur ihre Körperkanten durch die STD erkannt werden, sondern auch die Körpermitte mathematisch erfasst wird.
Infolge dieser Erkenntnise bilden wir die zwei weiteren Beobachtungen:

- (STD von AC_{1-63} < Threshold)$\wedge$!Outliers $\rightarrow$ Beobachtung AC_LOW
- (STD von AC_{1-63} > Threshold) $\vee$ Outliers $\rightarrow$ Beobachtung AC_HIGH

Auf Basis unserer Beobachtungen können nun anhand aller möglichen Permutationen von DC- und AC-Beobachtungen die Symbole unseres Eingabealphabets gebildet werden. Diese sind in Tabelle 1 dargestellt.

Alphabetsymbol	DC-Wert	Standartabweichung AC
Symbol 1	DC_BLACK	AC_LOW
Symbol 2	DC_BLACK	AC_HIGH
Symbol 3	DC_GREY	AC_LOW
Symbol 4	DC_GREY	AC_HIGH
Symbol 5	DC_WHITE	AC_LOW
Symbol 6	DC_WHITE	AC_HIGH

Tabelle 1: Eingabealphabet

2.6 Verfahren zur Vordergrundextraktion

Unser Verfahren zur Vordergrundextraktion basiert auf den zuvor genannten Algorithmen, der Bildpartitionierung in Blöcke und dem eingeführten Eingabealphabet. Jeder Block wird durch ein eigenes, individuelles HMM modelliert, da Blöcke unterschiedliche Übergangs-/Emissionswahrscheinlichkeiten aufgrund ihrer Lokalität aufweisen: ein Block an einer Tür ist öfter Veränderungen ausgesetzt, da erwartungsgemäß häufiger Personen durch diesen Bereich laufen, als ein Block in einem weniger frequentierten Bereich der eher dazu neigt kühl und dunkel zu sein.
In der ersten Phase ist ein Trainingsvideo zu verwenden, welches dazu dient, (individuelle) Trainingssequenzen für die Blöcke zu erstellen und im Folgenden diese dem Baum-Welch-Algorithmus zu übergeben, welcher die HMMs blocklokal trainiert. Da in einem Block hauptsächlich der Hintergrund zu beobachten ist, wird jeder Block nach der Lernphase seinen eigenen Hintergrund lernen und somit die korrespondierenden Emissionen als wahrscheinlicher bewerten.
In der zweiten Phase kann bereits ein Live-Video (in der Anwendung) eingesetzt werden, das HMM ist in dieser Phase zeitinvariant.

Für jeden Block wird die aktuelle Beobachtung gebildet. Die HMMs werden nun zur Vorhersage verwendet: Unter Verwendung des Forward- oder Viterbi-Algorithmus kann nun die nächste Emission prognostiziert werden. Diese wird mit der tatsächlich vorliegenden Emission (eine der 6 definierten Symbole unseres Eingabealphabetes) verglichen. Bei Gleichheit handelt es sich offensichtlich um den erlernten Hintergrund, bei Abweichungen ist ein Vordergrundobjekt in den Block eingetreten. Die aktuelle Emission wird an eine Warteschlange der k-letzten Emissionen angehangen und die Operationen wiederholt.

3 Implementierung und praktische Details

Dieses Kapitel gibt einen kurzen Überblick zu den von uns verwendeten Technologien, die wir zur Realisierung des theoretischen Entwurfs aus Kapitel 2 verwendet haben.

3.1 Videomaterial

Das uns zur Verfügung gestellte Bildmaterial besteht aus einzelnen Videosequenzen, wir können also nicht auf Live-Material arbeiten.

In der Praxis spielt dies jedoch keine große Rolle, da unter Linux der Zugriff auf eine Videodatei sich nicht wesentlich von dem Zugriff auf eine Kamera unterscheidet. Insgesamt standen uns sieben unterschiedliche Videosequenzen zur Verfügung, von denen sechs paarweise entstandene Aufnahmen sind. Das bedeutet, dass Kameraposition sowie Winkel zwischen den beiden Videosequenzen eines Paares nicht differiert, lediglich die aufgenommenen Szenen sind unterschiedlich. Diese Tatsache ist insofern hilfreich, als das wir die Möglichkeit haben, auf einer Videosequenz zu lernen und die erlernten Parameter später auf der anderen Videosequenz anzuwenden.
Der größte Nachteil der Verwendung von Videosequenzen ist, dass die einzelnen Sequenzen relativ kurz sind (Dauer übersteigt nicht 10 Minuten) und es somit nicht möglich ist, wirklich lange Lernphasen von zum Beispiel einigen Stunden zu realisieren. Der Vorteil ist hingegen, dass man Veränderungen an Programmbestandteilen und Parametern immer wieder an den selben Sequenzen ausprobieren kann und somit den entstehenden Effekt besser nachvollziehen kann.
Sämtliche Videosequenzen laufen mit 25 Frames pro Sekunde, was auf jeden Fall genug Daten zur Auswertung liefert. In der Praxis genügen vermutlich auch schon weniger Frames, da anhand der vorliegenden Videos deutlich wurde, dass sich die Personen im Bild nicht so schnell bewegen, dass sich innerhalb eines Bruchteils einer Sekunde die Szenerie stark verändert.

3.2 C++

Die Implementierung sowohl unseres Testbeds als auch die finale Implementierung als Komponente zur Vordergrundextraktion in einem vorgegebenen Framework fand in C++ statt. Für die Wahl von C++ gab es verschiedene Gründe, von denen vor allem die Performance, die Portabilität sowie das gute Angebot an zur Verfügung stehenden Bibliotheken im Vordergrund standen.

3.3 OpenCV

OpenCV[8] ist eine Bibliothek die eine Vielzahl unterschiedlicher Algorithmen für die Bildbearbeitung und somit letztlich für die Videobearbeitung zur Verfügung stellt. Hinzu kommt, dass auch Funktionen für das Lesen, Schreiben und Abspielen von Videodateien verschiedener Datentypen sowie von Kamera angeboten werden. Videos und Bilder werden in das OpenCV spezifische Format Mat eingelesen, auf dem viele Operationen möglich sind. Die von uns vorgestellte DCT ist Teil dieser Bibliothek und konnte von uns direkt auf den Mat-Objekten angewandt werden.
OpenCV bietet Interfaces für C, Python, Java und C++ und lässt sich auf vielen unterschiedlichen Plattformen einsetzen.

3.4 CvHMM

Aus der Menge der vorhandenen C++-HMM Bibliotheken fiel unsere Wahl auf CvHMM[9], vor allem da diese Bibliothek direkt auf dem in OpenCV[8] enthaltenen Standard-Videodatentyp Mat operiert und daher die Vermutung nahe liegt, dass weniger Performanceeinbußen vorliegen, da keine Konvertierung von Daten in andere Formate erfolgen muss.

4 Evaluation

In der letzten Phase unseres Projekts wurde uns ein Framework des SAFEST-Projekts[1] zur Verfügung gestellt, welches die Möglichkeit bietet, verschiedene Algorithmen miteinander zu vergleichen. Das Framework erhält als Eingabe eine Matrix, in der Vordergrundobjekte durch den zu testenden Algorithmus markiert sind, und berechnet im Folgenden basierend auf vordefinierten Abstandsparametern, die die Größe eines Menschen beschreiben, die Anzahl von Personen im Bild zu einem bestimmten Zeitpunkt.

Nach der Integration unseres Algorithmus in das Framework wird eine vergleichende Analyse mit der vorimplementierten, Histogramm-basierten Lösung durchgeführt.
Das ideale Verfahren wäre hier eigentlich, ein annotiertes Video zu verwenden, in dem für jedes Pixel bekannt ist, ob es zum Vorder- oder Hintergrund gehört. Da uns allerdings kein solches Video zur Verfügung stand, verwenden wir als Metrik den Abstand zur tatsächlichen Anzahl von Personen im Bild.
Die Werte der tatsächlichen Personenanzahl für diesen Vergleich sind uns seitens des SAFEST-Projekts[1] zur Verfügung gestellt worden und wurden manuell ermittelt. Das Ergebnis haben wir visualisiert, indem wir jeweils die Abweichung vom eigentlich erwarteten Wert als Graph dargestellt haben. Ein optimal funktionierender Algorithmus würde demnach als Graph im Diagramm ohne Abweichungen die Nulllinie verdecken. Es werden drei Videos von unterschiedlichen Umgebungen untersucht.
Im Folgenden werden die zwei Algorithmen wie folgt benannt:

Histogramm) die bereits implementierte Lösung, die mit Histogrammen arbeitet

HMM) unsere HMM-basierte Lösung, wobei zunächst mit einem Video gelernt wird und dann die Analyse auf einem zweitem Video der selben Szenerie statt findet

Für den HMM-basierten Algorithmus verwenden wir zum Decodieren den Viterbi-Algorithmus über die 3 letzten Beobachtungen und berücksichtigen vier Ausreißer per *outliers()*. Gelernt wurde jeweils über ein paarweise zugehöriges Video.
Weitere Vergleiche haben wir nicht angestellt, da das Histogramm-basierte Verfahren den anderen Verfahren entweder überlegen ist (OpenCV MOG und OpenCV MOG2[8]) oder aber das Verfahren den Hintergrund hart auf bestimmte Videos kodiert (MEAN) und sehr unflexibel ist, sobald der Hintergrund nicht mehr uniform ist.

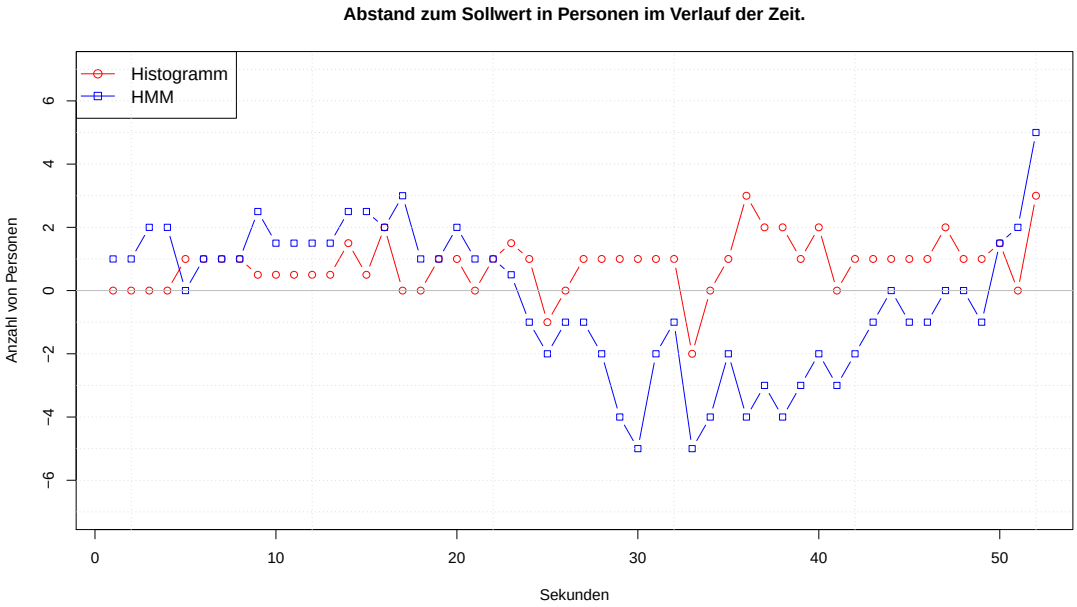

Abbildung 6: Innenhof: Histogramm vs. trainiertes HMM

4.1 Vergleich mit Histogramm-basierter Implementierung: Innenhof

Abbildung 6 zeigt auf, dass der Histogramm-basierte Algorithmus für dieses Video durchgehend bessere Ergebnisse liefert, als der HMM-basierte Algorithmus. Der HMM-Algorithmus ist im Durchschnitt 1-2 Personen weiter vom realen Wert entfernt als sein Konkurrent. Im Besonderen gilt dies für den zweiten Teil des Videos.

4.2 Vergleich mit Histogramm-basierter Implementierung: Tegel

Wie man der Abbildung 7 entnehmen kann, sind die Ergebnisse beider Algorithmen vom Verlauf her sehr ähnlich, allerdings weist der HMM-basierte Algorithmus gegenüber dem Histogramm-basierten kleine Vorteile auf. Zwischen 30 und 50 Sekunden hat er eine deutlich geringere Abweichung vom korrekten Wert. Dasselbe gilt für den Bereich zwischen ca. 65 und 80 Sekunden.

4.3 Vergleich mit Histogramm-basierter Implementierung: Eingang

In der Abbildung 8 erkennt man anhand der deutlich stärkeren Abweichung, dass beide Algorithmen Probleme mit der zuverlässigen Vordergrundextraktion haben.
Während in der ersten Hälfte des Videos der Histogramm-basierte Algorithmus noch Vorteile hat (relativ viele korrekte Anzeigen, der HMM-basierte Algorithmus erkennt

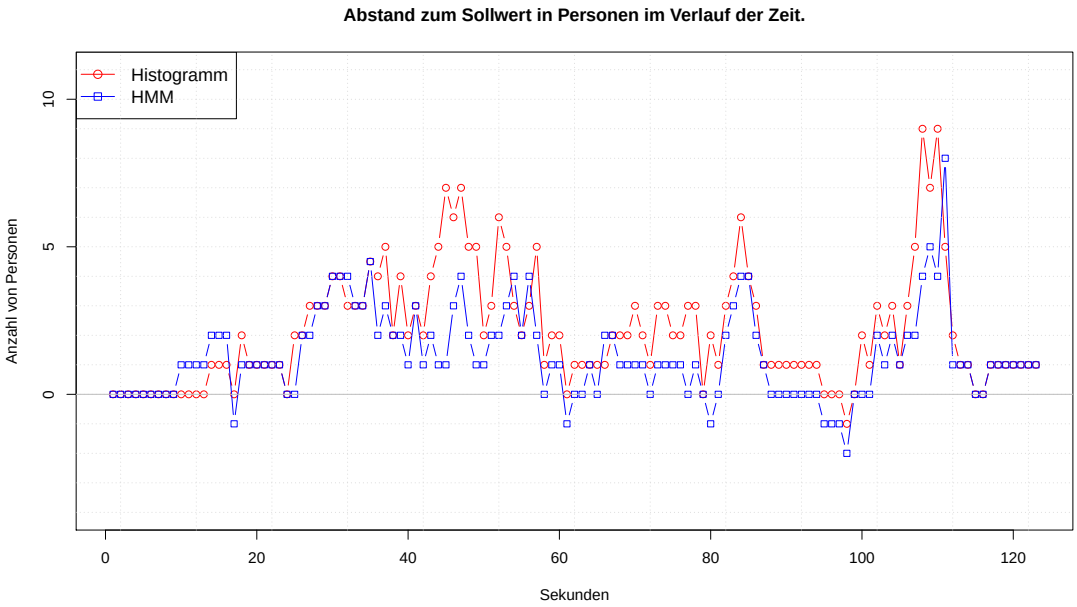

Abbildung 7: Tegel: Histogramm vs. trainiertes HMM

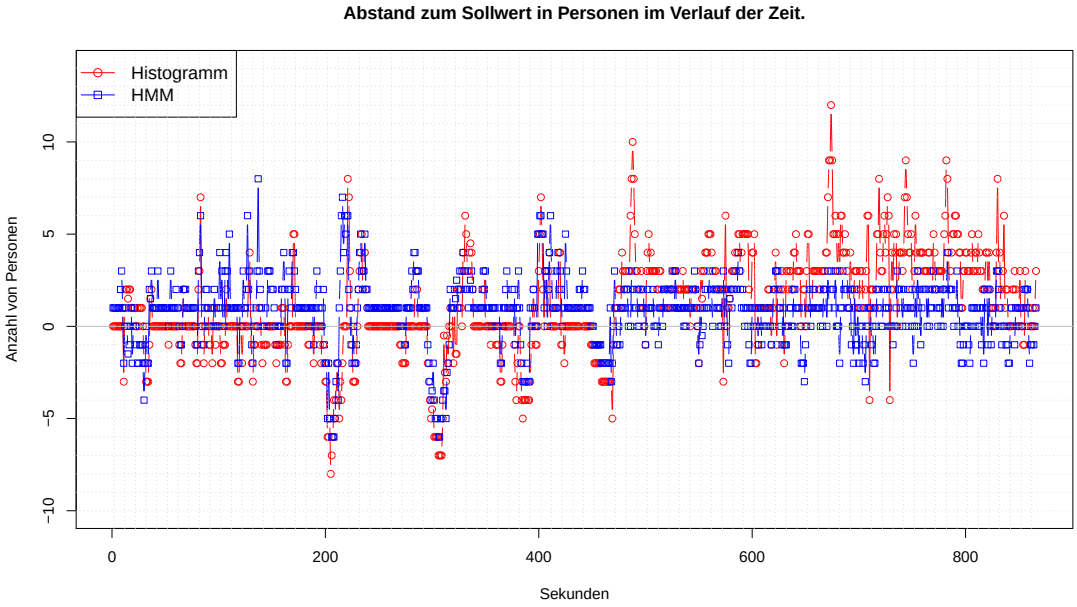

Abbildung 8: Eingang: Histogramm vs. trainiertes HMM (gesamt)

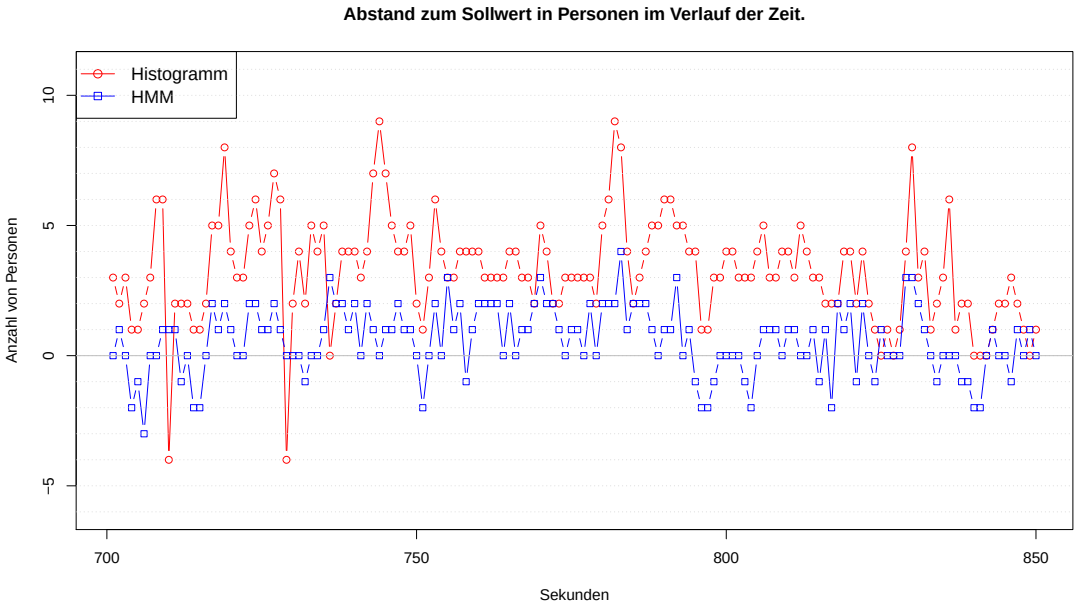

Abbildung 9: Eingang: Histogramm vs. trainiertes HMM, Ausschnitt Sekunde 700-850

meist eine Person zu viel), ist der HMM-basierte Algorithmus in der zweiten Hälfte des Videos (ab etwa Sekunde 450) deutlich besser. Wie der Grafik 9 zu entnehmen ist, existieren wesentlich weniger Ausreißer und das Ergebnis ist insgesamt auch viel dichter am korrekten Wert.

4.4 Diskussion

Aus den vorangehenden Grafiken kann man ablesen, dass beide Algorithmen grundsätzlich dazu geeignet sind, Vorder- von Hintergrund zu trennen. Sie besitzen jedoch Vor- und Nachteile in bestimmen Szenerien. Genau jene Unterschiede und Details müssen näher beschrieben und erklärt werden.

Wie aus Grafik 8 und 9 (Eingang) hervorgeht, besitzt der von uns entwickelte HMM-basierte Algorithmus besonders bei schwierigen, das heißt heterogenen, Hintergründen Vorteile bei der Erkennung. Trotz des hellen Hintergrunds, der viele Kanten und der häufigen Neukalibrierung der Kamera ist eine Erkennung mit kaum Ausreißern möglich. Dies liegt darin begründet, dass sich die Dynamik und die Veränderungen des Videobildes durch einen langen Lernprozess mitteln und des Weiteren Vordergrund nicht anhand von einfachen Grauwerten sondern anhand der komplexen (von uns eingeführten) Emissionen und den Übergängen zwischen ihnen erkannt wird.

Liegt jedoch ein homogener Hintergrund vor, so kann gesagt werden, dass beide Al-

gorithmen gut funktionieren, jedoch ist in dieser Situation der Histogramm-basierte Algorithmus besser. Das liegt daran, dass der Hintergrund einen bestimmten Grauwert besitzt und auch Personen in ein ganz spezielles Intervall der Grauwerte fallen. Demnach arbeitet hier der Histogramm-basierte Algorithmus sehr effizient, unser Algorithmus unterliegt in dieser Domäne.

Den größten Vorteil hat der Histogramm-basierte Algorithmus bei der Performance: während dieser mit 22-25 fps eine flüssige Videodarstellung ermöglicht, lief der HMM-basierte Algorithmus nur mit ca. 10-12 fps. Hier wäre also noch Potential zur Verbesserung bei unserem Algorithmus.
Leider hat sich im Laufe des Projekts gezeigt, dass CvHMM sehr ineffizient auf einzelne Daten zugreift und die Ursache hierfür ist. Dies konnten wir mit Hilfe von Callgrind (Teil von Valgrind) nachweisen. Ein weiterer großer Nachteil von CvHMM ist, dass nur eine begrenzte Anzahl von Algorithmen Angeboten wird. Es fehlt der Forward-Algorithmus, den wir mit Hinblick auf die Performance dem ineffizienteren Viterbi-Algorithmus vorziehen würden. Auch haben die Veränderungen an der Netzwerktopologie keine Veränderungen am Resultat ergeben, was wir auf die Implementierung des Baum-Welch-Algorithmus in CvHMM zurückführen. Somit konnte nicht überprüft werden, ob die von uns verwendeten zwei Zustände optimal waren.

Zusätzlich sei angemerkt, dass außer im Falle der Grafik 6 (Innenhof) die Verläufe beider Graphen sehr ähnlich sind. Dies legt nahe, dass die vorliegenden Fehler eine gemeinsame Ursache besitzen. Diese Ursache konnte von uns teilweise lokalisiert werden: das Clustering, welches markierten Vordergrund in die Anzahl von Personen übersetzt, ist oft zu großzügig. Wir konnten beobachten, dass bei den Tegel- und Eingang-Videos einzelne Personen in zwei oder mehr Cluster (Personen) geteilt werden.

Unseren Ergebnissen nach bewerten wir unser Algorithmus als einen allgemein einsatzfähigen Algorithmus, der in allen Szenerien gut genug ist (ein „allrounder“) der gegenüber dem spezialisierten Histogramm-basierten Ansatz in manchen Szenerien besser, in manchen schlechter abschneidet, aber stets gut genug ist.

5 Fazit und Ausblick

Wir sehen unsere Ziele der Anforderungsanalyse aus Kapitel 2.1 als erfüllt. Die in dem Abschnitt 4.4 angesprochenen Nachteile können sehr wahrscheinlich überwunden werden, wobei die Verwendung einer besseren Bibliothek für HMMs die erste wichtige Anpassung sein sollte. Ein möglicher Kandidat für eine bessere HMM-Bibliothek wäre dabei HMMlib[10], welches deutlich auf Performance hin optimiert wurde und des Weiteren den Forward-Algorithmus anbietet. Je nach eingesetzter Hardware kann auch Threading verwendet werden, um durch eine Einteilung des Bildframes in Bereiche und nichtsequentielle Berechnung die Abarbeitung zu beschleunigen. Zudem sollten algorithmische Anpassungen und theoretische Untersuchungen überprüfen, wie groß der Einfluss einer anderen HMM-Topologie wäre.

Wir werten das Ergebnis unseres Projekts als Erfolg. Wir haben nachgewiesen, dass es möglich ist, einen auf HMM-basierenden Algorithmus zur Trennung von Vorder- und

Hintergrund zu verwenden und haben mit diesen ähnlich gute Ergebnisse erzielt, wie mit den bereits implementierten Algorithmen. Bei heterogenen Hintergründen besitzt unser Konzept klare Vorteile gegenüber den Histogramm-basierten Algorithmus. Zusammenfassend erscheint der Einsatz eines HMM-basierten Vordergrundextraktors als sinnvoll.

Abbildungsverzeichnis

Literatur

[1] *Safest - social-area framework for early security triggers at airports*. Adresse: `http://safest.realmv6.org/`.

[2] N. Dalal und B. Triggs, "Histograms of oriented gradients for human detection", in *In CVPR*, 2005, S. 886–893.

[3] S. Dasgupta, "Learning mixtures of gaussians", in *FOCS*, 1999, S. 634–644.

[4] M. Stamp, *A revealing introduction to hidden markov models*, 2004.

[5] S. A. Khayam, *The discrete cosine transform (dct): theory and application". department of electrical & computing engineering*, 2003.

[6] M. Gales und S. Young, "The application of hidden markov models in speech recognition", *Found. Trends Signal Process.*, Bd. 1, Nr. 3, S. 195–304, Jan. 2007, ISSN: 1932-8346. DOI: `10.1561/2000000004`. Adresse: `http://dx.doi.org/10.1561/2000000004`.

[7] L. Yang, B. Widjaja und R. Prasad, "Application of hidden markov models for signature verification", *Pattern Recognition*, Bd. 28, Nr. 2, S. 161–170, 1995, ISSN: 0031-3203. DOI: `http://dx.doi.org/10.1016/0031-3203(94)00092-Z`. Adresse: `http://www.sciencedirect.com/science/article/pii/003132039400092Z`.

[8] *Opencv (open source computer vision)*. Adresse: `http://opencv.org/`.

[9] O. Sakhi, *Cvhmm - discrete hidden markov models based on opencv*. Adresse: `http://sourceforge.net/projects/cvhmm/,%20Sourceforge%20on%202012-06-10`.

[10] *Hmmlib*. Adresse: `http://www.cs.au.dk/~asand/?page_id=152`.

[11] M. Lamarre, *Tracking and Activity Classification in Video Surveillance Applications*. McGill University, 2002. Adresse: `http://digitool.library.mcgill.ca/webclient/StreamGate?folder_id=0&dvs=1393296165809~863`.