Selektive Videoaufzeichnung durch Klassifikation bewegter Regionen im komprimierten Bildraum

Selective video recording classifying moving regions in compressed domain

Uwe Rienäcker uwe.rienaecker@drivenet.de

In diesem Dokument wird ein Verfahren zur Klassifizierung bewegter Regionen für die selektive Videoaufzeichnung beschrieben. Die für die Bewegungserkennung und Extraktion der Objekteigenschaften benötigten Informationen werden ausschließlich aus dem komprimierten Bildraum gewonnen. Schwellwertoperationen auf der Differenz ausgewählter DCT Koeffizienten aufeinanderfolgender Frames liefern die Basis der Segmentation bewegter Regionen. Ein Regelwerk zur Clusterbildung verbindet zusammenhängende Bereiche des binären Differenzbildes und unterdrückt einzelne Blöcke ohne nachbarschaftliche Beziehungen. Für die Klassifikation der Regionen werden die jeweils kleinsten umschließenden rechteckigen Bereiche herangezogen. Aus bestimmten DCT Koeffizienten dieser Bereiche werden Merkmalsvektoren für die Klassifikation durch ein neuronales Netz extrahiert. Frames ohne erkennbare Bewegung oder ohne einer Objektklasse zugeordnete Regionen werden schon bei der Aufzeichnung verworfen. Die verbleibenden Informationen können für weitere Auswertungen verwendet oder einfach nur gespeichert werden. Die Vermeidung rechenzeitintensiver Operationen wie inverse DCT und Farbraumkonvertierungen ermöglicht eine Echtzeitverarbeitung der Videodaten mit geringen Anforderungen an die Hardware. Dieses wird durch experimentelle Ergebnisse bestätigt. Die spezielle Optimierung auf die Verarbeitung von Motion JPEG komprimierten Videodaten gab dem Projekt den Namen **M**otion **J**peg **CL**assification **S**ystem.

This paper describes a method to classify moving regions for selective video recording. All information needed for moving region segmentation and feature extraction are derived from compressed domain. The segmentation is based on thresholding the temporal difference between selected DCT coefficients of consecutive frames. A clustering rule set collects closed areas from the binary difference and suppresses single blocks without any neighbourhood relationship. For object classification the minimum bounding boxes of the segmented regions are used. An arithmetic scheme extracts feature vectors from selected DCT coefficients of these boxes. A neural network is used to classify the moving regions. Frames without detected movement or any classified object will be discarded during recording. The remaining information can be used for further evaluation or simply stored. Avoiding heavy computational operations like inverse DCT or color space conversions enables the method to run in real time on low hardware. This is validated by experimental results. Due to optimization for processing Motion JPEG compressed video the project is called **M**otion **J**peg **CL**assification **S**ystem.

1. Einleitung

Die visuelle Überwachung hat in den letzten Jahren immer mehr an Bedeutung nicht nur in der Sicherheitsbranche gewonnen. Immer mehr Technik zeichnet immer mehr Daten auf. Infolgedessen rücken automatisierte Analyseverfahren weiter in den Vordergrund der Forschung. Die Beobachtung monotoner Szenen auf der Suche nach außergewöhnlichen Vorkommnissen ist ein ermüdender Vorgang, insbesondere, wenn diese selten auftreten oder durch andere überlagert werden. Spätestens die zeitgleiche Überwachung mehrerer unabhängiger Szenen überfordert langfristig menschliche Fähigkeiten. Die Ereignisse von Interesse können simultan oder zeitlich versetzt mit oder ohne inhaltlichen Zusammenhang zueinander auftreten.

Eine präventive Beobachtung setzt zeitnahe Reaktionen auf außergewöhnliche Ereignisse voraus. Eine kontinuierliche Aufzeichnung ist nur für diagnostische Zwecke durch gezielte nachträgliche Auswertung geeignet.

Durch eine selektive Aufzeichnung können sowohl die Aufmerksamkeit schon zeitnah auf bestimmte Ereignisse gelenkt als auch die für eine nachträgliche Auswertung anfallenden Datenmengen erheblich reduziert werden. Eine reine bewegungsgesteuerte Aufzeichnung reicht dafür in den seltensten Fällen aus, da diese konsequenterweise alle Bewegungen des beobachteten Bereichs erfassen würde. Eine selektive Aufzeichnung muß daher schon eine qualitative Bewertung der Szenen vornehmen. Diese setzt die echtzeitfähige Verarbeitung der Videodaten voraus. Dabei stehen effektive Speicherung und schnelle Verarbeitung der Informationen oft im Widerspruch zueinander.

Unkomprimierte Rohdaten sind schneller zu verarbeiten, erfordern aber spezielle breitbandige Übertragungswege und Speichermedien mit hoher Kapazität. Daher wird nicht selten die Kompression in Form spezieller Hardware in die Aufzeichnungstechnik integriert. Die damit aufgezeichneten Daten können dann über vorhandene Infrastrukturen wie Netzwerke übertragen, zentral verarbeitet oder direkt ohne eine weitere Verarbeitung gespeichert werden.

Die automatisierte Auswertung wird nicht nur durch eine eventuell erforderliche Dekompression erschwert. Gängige Kompressionsverfahren sind verlustbehaftet. Sie sind auf maximale Datenreduktion bei minimalem Verlust an visueller Qualität optimiert. Die technische Qualität steht dabei weniger im Vordergrund. Gleiches gilt für oftmals verwendete "Bildverbesserungsverfahren". Diese dienen der optischen Verschönerung der aufgezeichneten Bilder, seltener der besseren technischen Verarbeitbarkeit, nicht zuletzt, weil in kostengünstiger Hardware gerade bei diesen Verfahren gespart wird. Für Auswertungsmethoden, die sich beispielsweise auf eine gleichbleibende Charakteristik oder die Linearität der Bildwandler verlassen, kann das zu einem Problem werden. Das betrifft bevorzugt Ansätze, die auf helligkeitsinvarianten Farbräumen basieren. Unabhängig davon liefert eine Dekompression der Bilddaten keinen Informationsgewinn. Im Gegenteil, ein Teil dieser müßte dann wieder aufwendig berechnet werden. Ziel muß es also sein, die vorhandenen Informationen effektiv zu verwerten statt aufwendig zu transformieren. Mit diesem Hintergrund wurde die vorgestellte Methode entwickelt, das heißt zum einen der vollständige Verzicht auf die Dekompression aber auch eine optimale Auswahl und Aufbereitung der zu verarbeitenden Informationen im gesamten Prozeß.

In diesem Dokument wird ein Ansatz zur Erkennung und Klassifikation von bewegten Regionen in komprimierten Videodaten, basierend auf ausgewählten DCT Koeffizienten, beschrieben. Speziell optimiert auf die Verarbeitung von Motion JPEG komprimierten Aufzeichnungen wird das Projekt **M**otion **J**peg **CL**assification **S**ystem genannt. Im folgenden Kapitel werden Betrachtungen über den Hintergrund der Entwicklung angestellt. Kapitel 3 beschreibt das Verfahren selbst. Im 4. Kapitel werden experimentelle Ergebnisse vorgestellt und im 5. Kapitel Schlußfolgerungen für zukünftige Entwicklungen gezogen.

2. Hintergrund

Eine Vielzahl mehr oder weniger umfangreicher Videoanalyseverfahren wurde in der Vergangenheit entwickelt und veröffentlicht. Die Erkennung und Auswertung von Bewegungen ist Basis der meisten Ansätze. Die Methoden sind jedoch unterschiedlich komplex.

Die Übersicht in Bild 1 beschreibt eine gängige Verfahrenskette. Die tatsächliche Reihenfolge der Prozeßschritte muß nicht immer mit der dargestellten identisch sein. Auch Interaktionen zwischen unmittelbar aufeinanderfolgenden Verarbeitungschritten wie beispielsweise die Korrektur von Objektzuordnung durch Konsistenz während der Spurverfolgung, beschrieben in [1], sind in diesem Schema nicht berücksichtigt. Die Betrachtungen in diesem Dokument beziehen sich auf die ersten drei, hervorgehobenen Bereiche. Andere vorgestellte Verfahren setzen auf vorhandenen Segmentationsmethoden eigene für das Tracking und die Klassifikation auf. Ansätze zur weitergehenden Interpretation sind dagegen seltener.

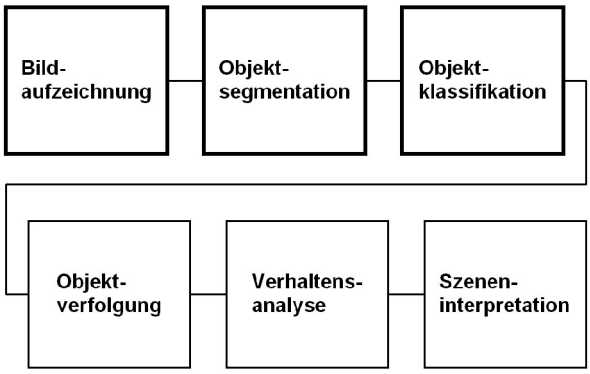

Bild 1: Übersicht Verarbeitungskette von Videoanalysesystemen

Das wohl bekannteste und umfassendste System zur visuellen Überwachung dürfte VSAM [1] sein. Die Architektur basiert auf der Vorverarbeitung der breitbandigen Bildinformationen unmittelbar am Ort der Aufzeichnung. Der Prozeß ist damit in eine zentrale und mehrere dezentrale Komponenten aufgeteilt. Zwischen den Einheiten werden symbolische Informationen und nur ausgewählte Bilddaten ausgetauscht. Zur gezielten Lenkung der Aufmerksamkeit auf auswählbare Ereignisse werden die Aufzeichnungseinheiten untereinander von zentraler Stelle aus koordiniert. Damit geht das Konzept weit über eine reine Videoaufzeichnung hinaus.

Ein nicht so umfangreiches ebenfalls interpretierendes Verfahren mit anderer Zielsetzung wird in [2] beschrieben. Durch Kombination von Hintergrundmodell und zeitlicher Differenz wird eine präzisere, auf die Erkennung von Hindernissen gezielte Segmentation vorgenommen. Im Unterschied zu [1] werden Methoden zur Unterdrückung von Schatten und Spuren bewegter Lichter eingesetzt. Durch die zeitliche Akkumulation von Informationen der Differenzbilder werden Aussagen über Bewegungsdichte und -fluß getroffen sowie plötzlich auftretende Hindernisse im beobachteten Bereich erkannt.

Ein großer Teil bisheriger Veröffentlichungen betrachtet einzelne Komponenten der oben beschriebenen Verfahrenskette. Schwerpunkte bilden dabei Segmentationsmethoden, Trackingverfahren sowie die Erkennung und Unterdrückung von stationären bzw. bewegten Schatten. Die meisten Verfahren verarbeiten Bildrohdaten auf Pixelebene. Für den Fall einer vorherigen Kompression müßte der Segmentation ein Dekompressionsschritt vorangestellt werden. Eine Echtzeitverarbeitung ist dann mit vertretbarem Aufwand nicht mehr erreichbar. Seltener sind Analyseverfahren für komprimierte Bilddaten zu finden. Die meisten von ihnen beziehen ihre Informationen für die Segmentation aus den Bewegungsvektoren der MPEG Codierung.

Prinzipiell gelten für die Verarbeitung komprimierter Bildinformationen in Form der DCT Koeffizienten unter Berücksichtigung eines gröberen Rasters die gleichen Gesetzmäßigkeiten wie für die Verarbeitung einzelner Bildpunkte. Darauf beruht das in [3] beschriebene Verfahren. Wahlweise werden Hintergrund und Differenzbild aus den Koeffizienten der Wavelet Transformation bzw. den DC Koeffizienten der DCT ermittelt und für ein bekanntes rekursives Hintergrundmodell [1] verwendet. Der Zeitgewinn wird durch eine geringere Auflösung und Präzision erkauft. So können die extrahierten Objekte nicht

wie in [6] beschrieben durch exakte Konturen voneinander und vom Hintergrund getrennt werden. Die ausschließliche Verwendung der DC Koeffizienten liefert ein sehr schnelles, gegenüber Rauschen unempfindliches Segmentationsverfahren. Nachteilig ist eine fehlende Robustheit gegenüber bewegten Schatten. Diese werden gemeinsam mit den bewegten Objekten extrahiert. Dadurch können unabhängige Objekte mitunter nicht mehr voneinander getrennt werden, was zu Klassifikationsfehlern führt.

Für die Erkennung bewegter oder stationärer Schatten werden sowohl geometrische als auch photometrische Eigenschaften herangezogen. Diese werden fast ausschließlich pixelbasiert ermittelt wie u.a. in [5], [6], [7].

Situationsbedingt sind nicht immer alle vorausgesetzten geometrischen Eigenschaften wie beispielsweise räumliche Lage der Schatten oder Verbindung zum Objekt gegeben. Die verwendeten photometrischen Eigenschaften sind dagegen universeller, setzen zum Teil aber verwertbare Farbinformationen voraus.

Interessant in diesem Zusammenhang ist das in [8] beschriebene Verfahren zur Gewinnung beleuchtungsinvarianter Bilder, da es prinzipiell unverändert auf aus den DC Koeffizienten der DCT berechneten Farbinformationen anwendbar ist und für diese identische aber nicht notwendig korrekte Ergebnisse liefert.

Geometrische Merkmale wie Position und Verlauf von Kanten lassen sich einfach aus Linearkombinationen der AC Koeffizienten berechnen.

Das im folgenden Kapitel beschriebene Verfahren nutzt bereits in den DCT Koeffizienten vorhandene Informationen über das Frequenzspektrum für eine teilweise Kompensation von bewegten Schatten aus.

3. Verfahrensbeschreibung

Auch wenn die Motion JPEG Kompression nicht die effektivste Variante der Bilddatenreduktion ist, hat sie den Vorteil, daß jeder Frame ein für sich eigenständiges Bild ist. Das Format ist einfach zu verarbeiten. Die Bilder lassen sich einzeln speichern und wieder zu einer Videosequenz zusammenfügen. Darüber hinaus ist dieses ein insbesondere bei Netzwerkkameras weit verbreitetes Format.

Mit dem Ziel der Anwendbarkeit für diese Kameras ist das Verfahren auf die Verarbeitung von Motion JPEG kodierten Videodaten optimiert. Experimentelle Ergebnisse werden zeigen, daß abhängig vom Anwendungsfall durch die selektive Aufzeichnung eine weit höhere Datenreduktion erzielbar ist als mit jedem noch so guten Kompressionsverfahren. Bild 2 zeigt schematisch den Ansatz, dessen Komponenten nachfolgend detaillierter beschrieben werden.

SOI-EOI Synchronisation

Der Vollständigkeit halber wird die Synchronisation der begrenzenden Marker SOI (Start Of Image) und EOI (End Of Image) erwähnt. Für eine unterbrechungsfreie Analyse müssen die Bilddatenblöcke unmittelbar aufeinander folgen. Trennzeichenfolgen, unvollständige oder fehlerhafte Bilder werden deshalb in diesem Prozeßschritt gefiltert.

DCT Extraktion

Aus den komprimierten Bilddaten werden nur diejenigen DCT Koeffizienten extrahiert und dequantisiert, die entsprechend dem konfigurierten Regelwerk für den weiteren Prozeß benötigt werden. Die Dequantisierung liefert keinen Informationsgewinn, dafür aber einheitliche Intensitätswerte unabhängig vom Kompressionsgrad.

Bewegungsextraktion

Die Bewegungsextraktion basiert auf der Differenz ausgewählter DCT Koeffizienten. Wahlweise kann die Differenz zum jeweils vorhergehenden Bild oder zu einem berechneten Hintergrund verarbeitet werden.

Die Differenz der Werte aufeinanderfolgender Bilder ist die einfachste und schnellste der in [4] diskutierten Methoden. Sie paßt sich schnell Hintergrundveränderungen an, erfaßt aber nicht die inneren Bereiche wenig texturierter Objekte. Darüber hinaus können schnelle Bewegungen Geisterspuren hinterlassen und zu langsame Objekte nicht oder unvollständig erfaßt werden. Alternativ für diesen Fall ist die Differenz zu einem Hintergrundbild vorgesehen. In diesem werden nur die nicht bewegten Regionen aktualisiert. Vordergrundbereiche, die sich über eine längere Zeit nicht verändern, werden als stationär angesehen und erst dann mit in den Hintergrund übernommen.

Bewegte Regionen werden durch Schwellwertoperationen segmentiert. Ein experimentell ermittelter Empfindlichkeitsparameter bestimmt zum einen die Verstärkung der Differenzsignale und zum anderen den Schwellwert für die Unterscheidung von Vorder- und Hintergrund. Durch ein Regelwerk arithmetischer und logischer Verknüpfungen der Differenzen einzelner Koeffizienten werden die Blöcke dem Vorder- oder Hintergrund zugeordnet.

In praktischen Tests, siehe Beispiele im Anhang, lieferte die Differenz mittelfrequenter DCT Koeffizienten aufeinanderfolgender Frames die präzisesten Ergebnisse. Das Schema reagiert somit auf texturelle Veränderungen. Im Vergleich zur Differenz der niederfrequenten oder zu der der DC Koeffizienten sind die Differenzen dieser robuster gegenüber bewegten Schatten, reagieren aber empfindlicher auf Rauschen. Das tritt wiederum nur bei wenig Licht auf, wenn bewegte Schatten keine Rolle mehr spielen, so daß die Wahl der Koeffizienten an die Tageszeit angepaßt werden kann. Die Tatsache, daß durch die Framedifferenz innere Bereiche schwach texturierter Objekte nicht erkannt werden, ist weniger von Bedeutung, wenn die Konturen ausreichend extrahiert werden. Parallel zur Bewegungsrichtung verlaufende Kanten größerer Objekte werden mit dieser Methode nicht erfaßt. In der Regel werden solche Objekte geteilt extrahiert, was in einem abschließenden Schritt der falsch negativ Reduzierung soweit wie möglich ausgeglichen wird.

Da das Verfahren von einem unbewegten Hintergrund ausgeht, ist es auf Aufzeichnungen stationärer Kameras beschränkt. Videosequenzen bewegter Kameras erfordern die Erkennung und Extraktion unabhängiger Bewegungen.

Zusammenfassung bewegter Regionen

Ein Regelwerk unter Berücksichtigung nachbarschaftlicher Beziehungen verbindet zusammenhängende Regionen des binären Differenzbildes und unterdrückt einzelne Blöcke ohne Nachbarschaft zu anderen dem Vordergrund zugeordneten Blöcken. Die kleinsten, die bewegten Regionen mit einem Abstand von einem Block umschließenden, rechteckigen Ausschnitte werden für die Objektklassifikation herangezogen. Ein zweiter höherer Schwellwert wird zur Erkennung von möglichen Hintergrundänderungen unter den selektierten Regionen verwendet. Ausschnitte, die keinen Block enthalten, dessen Diffenrenz auch diese Schwelle überschreitet, werden als potentielle Hintergrundkandidaten gekennzeichnet und von der direkten Klassifikation ausgeschlossen.

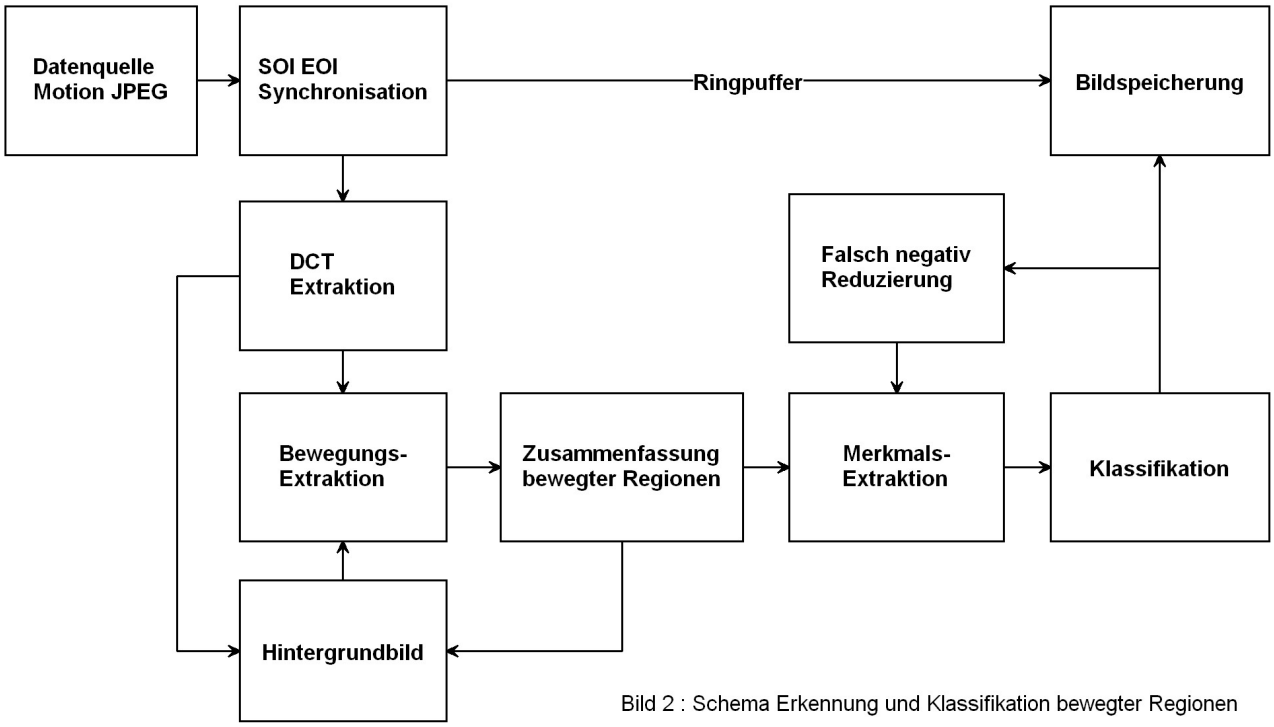

Bild 2 : Schema Erkennung und Klassifikation bewegter Regionen

Merkmalsextraktion Klassifikation

Durch arithmetische Verknüpfung ausgewählter niederfrequenter DCT Koeffizienten der rechteckigen Ausschnitte werden abstrakte Merkmalsvektoren generiert und auf eine einheitliche Länge skaliert. Diese enthalten die Verteilung von horizontalen und vertikalen Kanten innerhalb des rechteckigen Bereichs in kodierter Form. Die in den Bildausschnitten enthaltenen Informationen müssen soweit komprimiert werden, daß eine Generalisierung des Klassifikators schon mit kleinen Trainingsmustermengen erreichbar ist, die Klassen selbst aber unterscheidbar bleiben. Gemeinsam mit Ausdehnung und Seitenverhältnis der Regionen bilden diese Vektoren die Merkmale für die Klassifizierung durch ein neuronales Netz. Die im vorhergehenden Prozeßschritt als potentielle Hintergrundkandidaten markierten Regionen werden nicht direkt klassifiziert.

Die Objekte werden durch ein hybrides System konkurrierender dreischichtiger Perceptrons, trainiert nach dem Error Backpropagation Algorithmus, unterschieden. Dabei wird jeder zu erkennenden Klasse ein Ausgangsneuron zugeordnet, dessen Wert bei Klassenzugehörigkeit positiv andernfalls negativ sein soll. Für keiner Klasse zuzuordnende Objekte bleiben alle Ausgänge negativ. Die Klassenzuordnung erfolgt anhand des Ausgangs mit dem höchsten über einer Schwelle liegenden Wert. Unterschreiten die Werte aller Ausgänge diese Schwelle, wird das Objekt keiner Klasse zugeordnet.

Die untereinander konkurrierenden Netze des Klassifikators sind in Gruppen mit verschiedenen Transformationsfunktionen an Ein- und Ausgang zusammengefaßt und jedes einzelne für sich individuell konfiguriert.

In Bild 3 ist das System schematisch dargestellt. Der Einfluß jeder Gruppe auf das Gesamtergebnis und der jedes einzelnen Netzes innerhalb der Gruppe wird durch den Erfolg während des Trainings bestimmt.

Trainiert wird das System anfangs mit einer kleinen Menge von Hand markierter Beispiele, von denen ein Teil als Trainings- und ein Teil als Referenzmenge dient. Das Training ist beendet, wenn sich der Fehler auf der Referenzmenge nach einer festgelegten Zahl Trainingszyklen nicht mehr verringert. Im weiteren Verlauf wird die Trainingsmenge durch falsch klassifizierte Beispiele ergänzt bis entweder die gewünschte Klassifikationsgüte erreicht ist oder sich keine Verbesserung mehr einstellt.

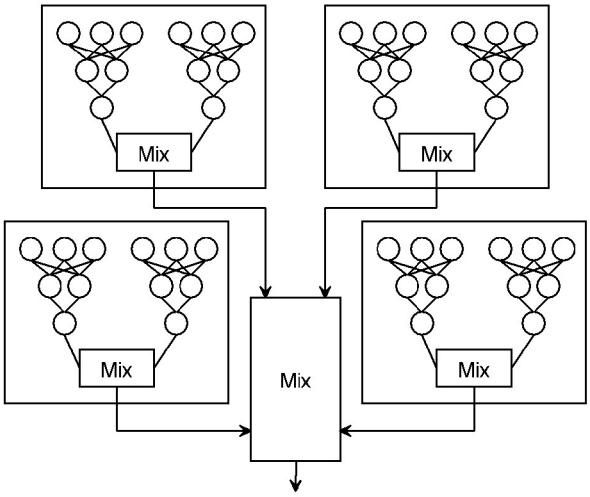

Bild 3 : Schema hybrides System konkurrierender Netze

Falsch negativ Reduzierung

Ziel ist es, die Objekte lückenlos zu erkennen. Das ist in den seltensten Fällen möglich. Ursachen dafür gibt es mehrere. Einerseits kann die Generalisierungsfähigkeit des Klassifikators noch nicht ausreichen. Dann empfiehlt sich ein Nachtraining mit den nicht erkannten Beispielen. Andererseits kommen ungewollte Teilungen von Objekten durch unzureichende Segmentation in Frage. Diese sind oft als Teilbereiche mit einer bestimmten räumlichen Lage zueinander zu erkennen. Werden Regionen ohne Zuordnung zu einer Klasse mit einer für eine ungewollte Teilung

typischen räumlichen Lage zueinander gefunden, werden diese zu einer zusammengefaßt und erneut einer Klassifikation unterzogen. Ist dadurch eine Klassenzuordnung möglich, werden die bisherigen Teilregionen durch die zusammengefaßte ersetzt. An diesem Prozeß werden auch die als potentielle Hintergrundkandidaten gekennzeichneten Bereiche beteiligt. Sie könnten in Folge einer ungewollten Teilung entstanden sein.

Bildspeicherung

Für die Betrachtung oder weitere Verarbeitung werden die positiv klassifizierten Einzelbilder, bei denen wenigstens eine bewegte Region gefunden und einer Klasse zugeordnet werden konnte, mit Zeitpunkt der Aufzeichnung, laufender Nummer und den Koordinaten der extrahierten Bereiche sowie deren Klassenzugehörigkeit gespeichert. Um dabei Zeit zu sparen, werden die original aufgezeichneten Bilddaten direkt aus einem Ringpuffer übernommen.

4. Experimentelle Ergebnisse

Im Rahmen der Tests bestand die Aufgabe darin, sich bewegende Personen zu erkennen und von allen anderen Bewegungen wie Fahrzeuge, Tiere oder bewegte Bäume zu unterscheiden. Die Aufzeichnung erfolgte mit einer Axis 2100 Netzwerkkamera zu verschiedenen Tageszeiten in einem Wohngebiet mit wenig bis mittlerem Verkehr mit einer Bildfolge von 15/Sekunde. Für das Training wurden ca. 1000 automatisch extrahierte Ausschnitte von 55 verschiedenen Ereignissen manuell gekennzeichnet. In 5 weiteren Etappen wurde die Trainingsmenge um weitere, bis dahin nicht korrekt zugeordnete Beispiele ergänzt. Ab einem Umfang von 4000 Trainingsbeispielen stellte sich keine weitere Verbesserung der Erkennungsleistung ein.

Das Ergebnis des kombinierten Trainings wurde für die Aufzeichnung von 10 Serien an mehreren Tagen zu unterschiedlichen Tageszeiten über einen Gesamtzeitraum von 21 Stunden verwendet. Der zeitliche Anteil der Bewegungen betrug knapp 3 Stunden. Die Hälfte davon wurde durch Bewegungen von Bäumen an einem windigen Tag verursacht. Für die Bewertung der Ergebnisse wurden die falsch positiven Zuordnungen durch Sichtung der Aufzeichnungen ermittelt, die falsch negativen automatisch durch entstandene Lücken in der laufenden Numerierung innerhalb eines Ereignisses identifiziert. Die folgende Tabelle 1 enthält eine Übersicht der Ergebnisse in absoluten Werten.

Serie	*Bewegungen*	*positiv*	*negativ*	*f positiv*	*f negativ*
1	26102	3387	22715	202	799
2	49158	3113	46045	97	504
3	36065	2886	33179	46	268
4	8322	3254	5068	43	382
5	8089	3131	4958	21	295
6	1471	684	787	5	64
7	1756	741	1015	6	163
8	7412	2771	4641	29	379
9	6073	2819	3254	57	220
10	8397	2624	5773	134	264
gesamt	152845	25410	127435	640	3338

Tabelle 1 Klassifikationsergebnisse realer Aufzeichnungen Werte absolut

In der folgenden Tabelle 2 sind die prozentualen Anteile der falsch zugeordneten Bilder dokumentiert. Dabei wurde der Anteil der falsch positiven Zuordnungen als Quotient der Anzahl falsch positiver und der Anzahl tatsächlich negativer Einzelbilder berechnet. Der Anteil falsch negativer Zuordnungen ergibt sich als Quotient der Anzahl falsch negativer und der Anzahl tatsächlich positiver Einzelbilder. Die Spalte Verhältnis gibt die Reduktion der Daten aller Bewegungen durch die Einschränkung auf die positiv bewerteten Bilder wieder. Unter Einbeziehung des zeitlichen Anteils der Bewegungen an der gesamten Aufzeichnungszeit ergibt sich in diesem Beispiel eine Reduktion der Datenmenge auf 1/50 des ursprünglichen Volumens.

Serie	*Bewegungen*	*f positiv %*	*f negativ %*	*Verhältnis*
1	26102	0,89	23,6	0,106
2	49158	0,21	16,2	0,055
3	36065	0,14	9,3	0,074
4	8322	0,85	11,7	0,350
5	8089	0,42	9,4	0,353
6	1471	0,64	9,4	0,425
7	1756	0,59	22,0	0,332
8	7412	0,62	13,7	0,327
9	6073	1,75	16,1	0,437
10	8397	1,05	7,8	0,297
gesamt	152845	0,50	13,1	0,149

Tabelle 2 Klassifikationsergebnisse realer Aufzeichnungen Werte relativ

Die Ergebnisse der 8. Serie, deren Einzelergebnis nahe dem Durchschnitt liegt, wurden genauer analysiert. Eine Betrachtung der falsch positiven Zuordnungen fand wegen ihres geringen Anteils nicht statt. Innerhalb der 8. Serie wurden 38 unabhängige Ereignisse aufgezeichnet. 87 von den 379 falsch negativen Bewertungen hinterließen Lücken von jeweils nur einem Einzelbild, was bei der visuellen Beurteilung in den wenigsten Fällen aufgefallen ist. Die verbleibenden 292 Fehlentscheidungen verteilten sich auf insgesamt 81 Lücken von 2 oder mehr Frames. Eine gegebenenfalls angeschlossene Objektverfolgung hätte in 5 Fällen durch eine zu große Lücke die Spur verlieren können. Tabelle 3 faßt diese Daten zusammen.

Ereignisse	*Lücken=1 F*	*Lücken>1 F*	*verl. Spur*
38	87	81	5

Tabelle 3 Detaillierte Betrachtung Serie 8 Teil 1, Aufteilung der falsch negativen Klassifikationen

Für die in dieser Aufzeichnung entstandenen größeren Lücken von 2 oder mehr Einzelbildern wurde die Verteilung der 3 häufigsten Fehlerursachen ermittelt.

Eine Fehlklassifikation wurde angenommen, wenn die beiden anderen Ursachen ausgeschlossen werden konnten.

Überdeckung	*wenig Bewegung*	*Fehlklassifikation*
40	27	14

Tabelle 4 Detaillierte Betrachtung Serie 8 Teil 2, Verteilung der Ursachen für die entstandenen Lücken von mehr als 1 Frame

Das Verfahren ist gemischt in ANSI C (Bildaufbereitung) und C++ (Neurosimulator) implementiert, lauffähig sowohl auf Windows als auch UNIX/LINUX Plattformen. Zur Aufbereitung der JPEG komprimierten Bilddaten wurde die Bibliothek der Independend JPEG Group um einen Ringpuffer ergänzt. Die Kernroutinen zur Bewegungs- und Merkmalsextraktion wurden speziell für dieses Verfahren implementiert und von Hand laufzeitoptimiert. Die Tests, ausgenommen das Training, wurden auf einem Pentium 133

PC unter LINUX mit der maximalen durch die Kamera erreichbaren Bildfolge von 15/Sekunde durchgeführt. Bei einer dabei erreichten Systemauslastung von ca. 75% können somit auf diesem System bis zu 20 Bilder der Dimension 320x240 je Sekunde verarbeitet werden.

5. Schlußfolgerungen

Die Ergebnisse zeigen, daß mit geringem Aufwand ohne Verwendung von Kontextinformationen beachtliche Resultate bei der selektiven Videoaufzeichnung erzielbar sind. So ließen sich im demonstrierten Beispiel der Inhalt von 20 Stunden Aufzeichnung auf die Ereignisse von weniger als ½ Stunde reduzieren, ohne wesentliche Informationen zu verlieren. Für die Zukunft stehen Tests mit weiteren Datenquellen aus. Der konsequente Verzicht auf zeitaufwendige Operationen macht es zu einem der schnellsten Verfahren seiner Art. So können auf einem heutigen Standard PC mehr als 10 der im Test verwendeten Kameras gleichzeitig bedient werden. Die Optimierung auf die Verarbeitung von Motion JPEG Videodaten ermöglicht die Nutzung kostengünstiger Hardware, die einfach in vorhandene Infrastrukturen integriert werden kann.

Zukünftige Entwicklungen werden sich auf eine weitere Reduktion falsch negativer Zuordnungen, z.B. als Folge von Überdeckungen, durch Hinzunahme bisher nicht verwendeter Kontextinformationen konzentrieren. So könnte die Klassifikation durch eine Komponente der zeitlichen Kontinuität unterstützt werden. Erreicht wird dies z.B. durch die Integration eines Trackingverfahrens ähnlich der in [9] oder [10] beschriebenen.

Ein weiteres Potential besteht in der Untersuchung von Verfahren zur Gewinnung beleuchtungsinvarianter Bilder zur Unterdrückung von Schatten aus dem komprimierten Bildraum. Beispielsweise liefert der in [8] beschriebene Ansatz für eine handelsübliche Videokamera stabile und korrekte Ergebnisse, für die in diesem Projekt verwendete Netzwerkkamera sowie für viele im öffentlichen Raum eingesetzte Web- Kameras ebenfalls stabile aber falsche Ergebnisse. Dabei sind die ermittelten invarianten Transformationsrichtungen für ein Bild identisch, unabhängig davon, ob sie auf Pixelebene oder aus den DC Koeffizienten der DCT ermittelt wurden. In beiden Fällen liefern sie keine schattenfreien Bilder, da eine tatsächlich invariante Transformation entweder schwer zu ermitteln ist oder in eine andere als die ermittelte Richtung verläuft. Eine Lösung könnte u.a. in der Kompensation von Nichtlinearitäten der Aufzeichnungstechnik liegen.

Im Zusammenhang mit einer aktuellen Entwicklung zur Verbesserung der Objektextraktion durch die Verwendung eines neuen Hintergrundmodells gewinnt die Kompensation bewegter Schatten erneut an Bedeutung. So konnte durch die Anwendung eines approximativen Median Filter zur Anpassung der bewegten Regionen im Hintergrundmodell die Präzision bei der Segmentation insbesondere unter schlechten Lichtverhältnissen deutlich verbessert werden. Durch diese Veränderung konnten gleiche oder bessere Klassifikationsergebnisse mit nur noch einem Bruchteil der Trainingsmenge erreicht werden. In gleichem Maß erhöhte sich die Empfindlichkeit gegenüber bewegten Schatten, wodurch die Präzision bei direktem Sonnenlicht wiederum beeinträchtigt wird. Ausführliche Tests stehen noch aus und werden Gegenstand weiterer Betrachtungen sein.

Literatur

[1] Collins, R.T., Lipton, A.J. & Kanade, T. (1999) „A system for video surveillance and monitoring“.

[2] Naohiro Amamoto, Akihiro Fujii, „Detecting Obstructions and Tracking Moving Objects by Image Processing Technique“, Electronics and Communications in Japan, Part 3, Vol. 82, No. 11, 1999

[3] B. Ugur Toreyin, A. Enis Cetin, Anil Aksay, M. Bilgay Akhan, „Moving Region Detection in Compresed Video“, Aykanat et al. (Eds.): ISCIS 2004, LNCS 3280, pp. 381–390, 2004.

[4] Sen-Ching S. Cheung, Chandrika Kamath, „Robust techniques for background subtraction in urban traffc video“, Video Communications and Image Processing, SPIE Electronic Imaging, San Jose, January 2004, UCRL-JC-153846-ABS, UCRL-CONF-200706

[5] Alessandro Bevilacqua, „Effective Shadow Detection in Traffic Monitoring Applications“, WSCG 2003, Vol. 11, no. 1

[6] A. Bevilacqua, M. Roffilli,“Robust denoising and moving shadows detection in traffic scenes“, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Dec 2001

[7] Andrea Cavallaro, Elena Salvador, Touradj Ebrahimi, „Detecting Shadows In Image Sequences“

[8] Graham D. Finlayson, Mark S. Drew, Cheng Lu, „Intrinsic Images by Entropy Minimization“, ECCV, Prague, 2004

[9] Jonathan Owens, Andrew Hunter, Eric Fletcher, „A Fast Model-Free Morphology-Based Object Tracking Algorithm“, British Machine Vision Conference, BMVC2002

[10] Beleznai C., Schloegl T., Wachmann B., Bischof H., Kropatsch W., „Tracking multiple objects in complex scenes“, Proc. of 26th Workshop of the Austrian Association for Pattern Recognition, F. Leberl and F. Fraundorfer (eds), vol. 160, pp. 175 – 182, Austrian Computer Society, 2002

Anhang

Die Bilder 4 bis 9 demonstrieren beispielhaft die durchschnittlich höhere Präzision der Bewegungsextraktion durch Framedifferenz der mittelfrequenten DCT Koeffizienten gegenüber der der niederfrequenten. Für 3 Beispiele sind jeweils das kontrastoptimierte Differenzbild auf Pixelebene links, das Differenzbild der DCT Koeffizienten mitte und die durch Clusterbildung zusammengefaßten Bereiche rechts dargestellt. Die Rechtecke links markieren die für die Klassifikation herangezogenen Regionen. Das jeweils erste Bild der 3 Pärchen ist das Ergebnis unter Verwendung der 3 niederfrequentesten AC Koeffizienten. Das zweite entsteht bei Verwendung von 3 Koeffizienten aus dem mittleren Frequenzbereich.

Im ersten Beispiel wird die Person durch Verwendung der mittelfrequenten Koeffizienten erst vollständig erfaßt. Die im zweiten Beispiel die Personen verbindenden Schatten werden bei gleichem Schwellwert teilweise nicht extrahiert und dadurch die Personen getrennt erkannt. Nur im dritten Beispiel gehen Informationen verloren.

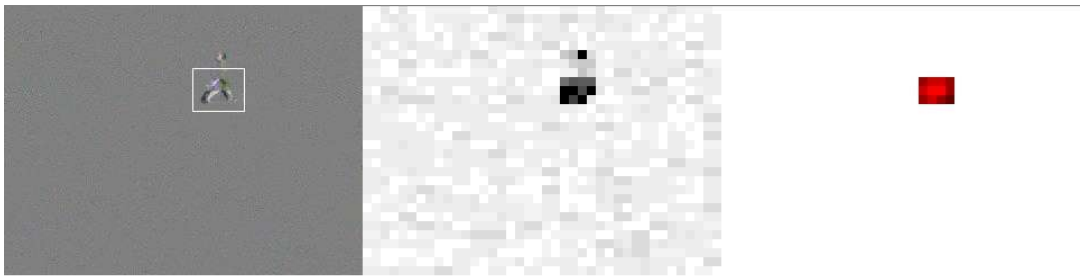

Bild 4 : Beispiel 1 Verwendung niederfrequenter Koeffizienten Differenzbild und extrahierter Bereich links, DCT Differenz mitte, zusammenhängende Regionen rechts

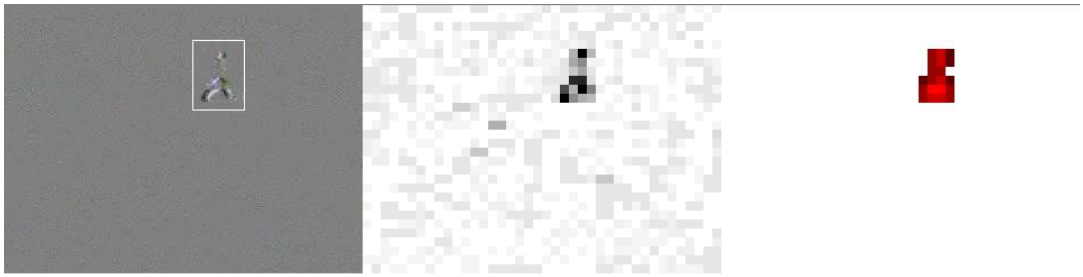

Bild 5 : Beispiel 1 Verwendung mittelfrequenter Koeffizienten Differenzbild und extrahierter Bereich links, DCT Differenz mitte, zusammenhängende Regionen rechts

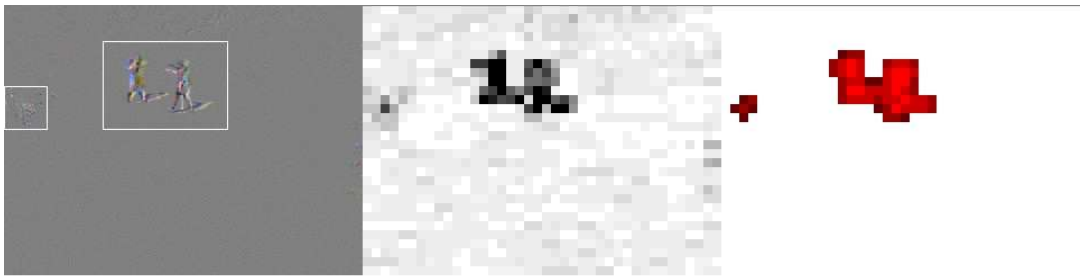

Bild 6 : Beispiel 2 Verwendung niederfrequenter Koeffizienten Differenzbild und extrahierter Bereich links, DCT Differenz mitte, zusammenhängende Regionen rechts

Bild 7 : Beispiel 2 Verwendung mittelfrequenter Koeffizienten Differenzbild und extrahierter Bereich links, DCT Differenz mitte, zusammenhängende Regionen rechts

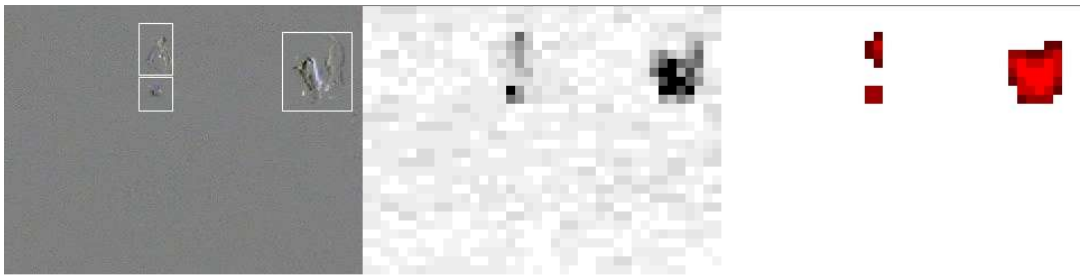

Bild 8 : Beispiel 3 Verwendung niederfrequenter Koeffizienten Differenzbild und extrahierter Bereich links, DCT Differenz mitte, zusammenhängende Regionen rechts

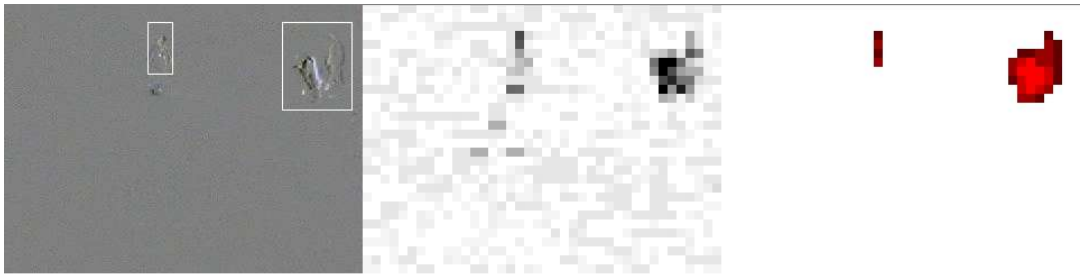

Bild 9 : Beispiel 3 Verwendung mittelfrequenter Koeffizienten Differenzbild und extrahierter Bereich links, DCT Differenz mitte, zusammenhängende Regionen rechts