Hochschule **Rosenheim**
University of Applied Sciences

Erkennung von audiovisuellen Events in Überwachungskameraaufnahmen

Manuel Huber
Fakultät für Informatik

SS 2018

Im Bereich der Videoüberwachung ist automatische Szenenanalyse und Eventerkennung eine Möglichkeit große Mengen von Aufnahmen ohne menschliche Arbeit auszuwerten. Hierbei wird immer stärker nicht nur anhand der Videodaten analysiert, sondern auch anderer Modalitäten, wie zum Beispiel Audio. In dieser Veröffentlichung wird die sogenannte AVC-Matrix-Methode vorgestellt, welche zunächst Video- und Audiosignale unabhängig voneinander analysiert. Sie werden mit einem Standardverfahren verarbeitet, welches Vord- und Hintergrund-Aktivität unterscheiden kann und gleichzeitig permanente Veränderungen des Hintergrundes erkennt. Anschließend werden Audio- und Videovordergrund anhand ihrer zeitlichen Synchronität in der Audio-Video-Gleichzeitigkeits-Matrix (Audi-Video-Concurrency Matrix, AVC Matrix) kombiniert und können somit als audiovisuellen Events erkannt werden. Hierfür wird die Tatsache ausgenutzt, dass die zu erkennenden Events gleichzeitig sowohl akustische wie auch visuelle Aktivität zeigen. Experimentelle Ergebnisse zeigen bessere Ergebnisse, wie eine Analyse mit nur einer einzelnen Modalität oder beider Modalitäten ohne Einbezug der zeitliche Synchronität.

Inhaltsverzeichnis

1 Einleitung

Die steigenden Sicherheitsanforderungen von öffentliche Plätzen, kritischer Infrastruktur oder privater Grundstücken fordern oft kontinuierliche Videoüberwachung vieler Lokationen. Diese gewaltige Menge an Daten muss in Echtzeit ausgewertet werden, um sofort auf Gefahren reagieren zu können, welches in der Vergangenheit nur durch menschliche Operatoren möglich war. Eine Automatisierung dieser Aufgabe verringert die Kosten und steigert gleichzeitig die Zuverlässigkeit, weshalb Videosequenzanalyse und Mustererkennung immer mehr an Bedeutung gewinnen. Das Ziel ist es komplexe Aktivitäten und Akteure in einer Videoaufnahme zu erkennen und kategorisieren.

Solche Analysen sind oft hierarchisch aufgebaut, wobei zuerst Analysen auf niedriger Ebene durchgeführt werden, wie zum Beispiel Vorder- und Hintergrundanalyse [Sta98]. Hierbei werden die erwarteten Elemente des Bildes (der Hintergrund) von den unerwarteten (Vordergrund) getrennt.

Viele Systeme, welche menschliche Aktivität erkennen, arbeiten ausschließlich mit visuellen Daten, aber andere Modalitäten wie Audio sind oft zusätzlich vorhanden und können genutzt werden um Aktivitätsmuster genauer zu erkennen.

2 AVC Matrix Analyse

2.1 Übersicht

Das System soll audiovisuelle Aufnahmen von Überwachungskameras analysieren um Ereignisse zu erkennen und zu kategorisieren. Im Gegensatz zu zum Beispiel dem in 3.1 vorgestellten System, soll dieses System mit lediglich einer Kamera und einem Mikrophon arbeiten. Hierfür werden die Audio- und Videodaten zunächst einzeln analysiert und anschließend in einer Audio-Video-Concurrency AVC-Matrix kombiniert um Ereignisse zu erkennen. Diese Methode basiert darauf, dass Ereignisse sowohl visuell als auch akustisch erkennbar sind. Ziel ist es eine bessere Klassifizierung von Ereignissen zu erzielen wie mit Audio oder Video alleine möglich ist.

Die AVC Matrix Analyse besteht aus mehreren Einzelschritten, welche in Abbildung 1 zu sehen sind.

Die Videoaufnahmen werden zunächst auf Pixelebene mit einem adaptiven Verfahren aus mehreren Gausmodellen analysiert um den visuellen Vordergrund vom Hintergrund zu unterscheiden [Sta98]. Anschließend werden die Vordergrunddaten in ein Farben-Histogramm umgewandelt und von einem zweiten Modell analysiert, welches verschiedene Ereignisse kategorisieren kann.

Ein Audio-Hintergrund-Modell wird verwendet um unerwartete Geräusche zu erkennen und einen akustischen Vordergrund herauszufiltern. Hierfür werden die Audio Aufnahmen in Frequenz-Balken unterteilt. Anschließend wird ein adaptives Verfahren aus mehreren Gausmodellen verwendet um ein Modell für jeden Balken zu trainieren, welches es erlaubt Hintergrundgeräusche zu markieren und somit unerwartete Geräusche (Vordergrundgeräusche) zu erkennen.

Audio- und Videoanalyse laufen parallel und erzeugen für jeden Zeitschritt t einen separaten Audio- und Video-Vordergrund. Diese Daten werden anschließend zu einer sogenannten AVC-Matrix kombiniert um die Gleichzeitigkeit der Audio- und Video-Vordergrund-Muster zu analysieren.

In psychologischen Studien wurde festgestellt, dass Menschen davon ausgehen, dass zeitlich korrelierende Audio- und Video-Ereignisse auch eine kausale Verbindung haben [Nie02]. Deshalb werden die Vordergrundaufnahmen von Audio und Video in der AVC Matrix kombiniert um anschließend die Gleichzeitigkeit von Audio- und Video-Vordergrundereignissen zu quantifizieren. Ein AVC Matrix Eintrag $[i, j]$ repräsentiert die Vordergrundgeräusche des $i$ten Balken und dem Erscheinen eines unerwarteten Farbspektrums des $j$ten Balken des Video-Vordergrund-Histogramms. Diese Methode erlaubt es die Gleichzeitigkeit der Ereignisse über einen längeren Zeitraum zu bewerten und die somit die wahrscheinlichsten Audio-Video-Ereignis-Kombinationen zu erkennen.

Diese Kombination der Modalitäten Audio und Video erzielt eine höhere Erfolgsrate bei Gruppierung und Klassifizierung als eine Analyse mit nur einer der Modalitäten.

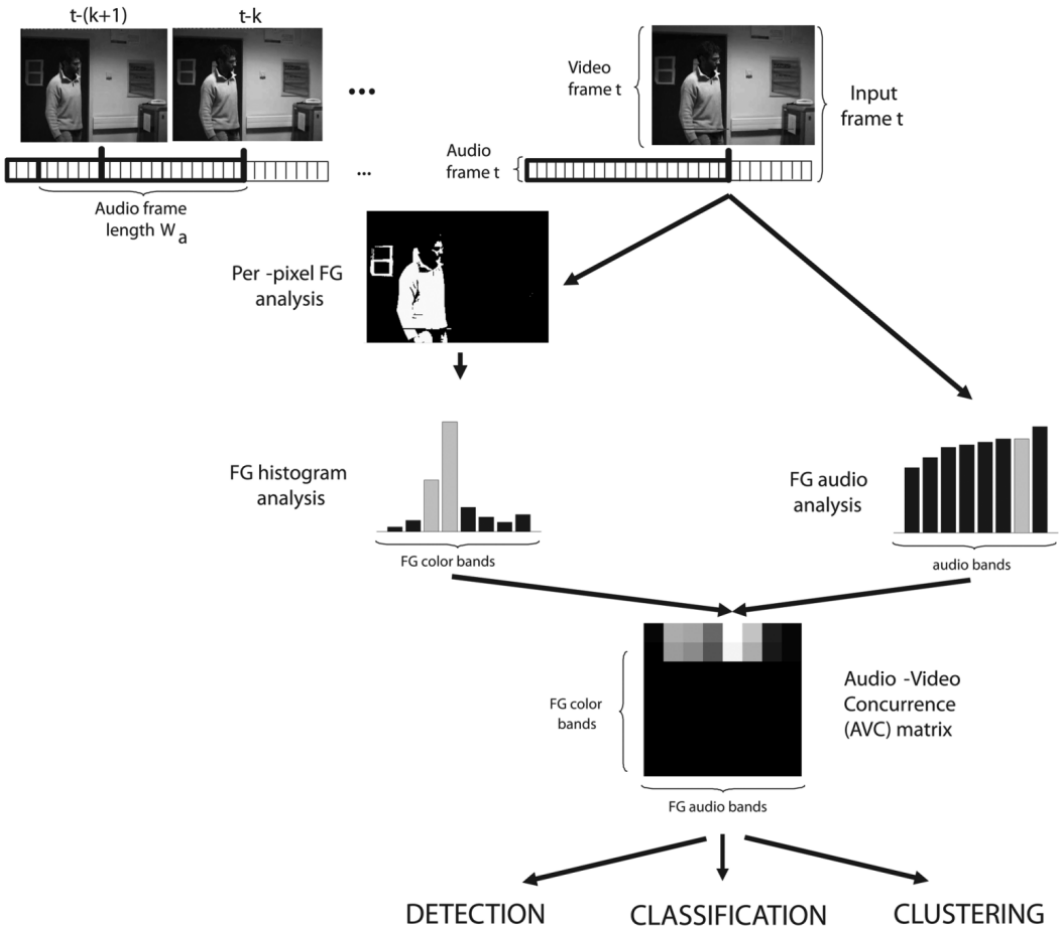

Abbildung 1: Überblick über den Ablauf der AVC Matrix Analyse

2.2 Zeitadaptive Kombination von Gaußverteilungen

Die Zeitadaptive Kombination von Gaußverteilungen (Time-Adaptive Mixture of Gaussians, kurz TAPPMOG) ist eine Methode zur Erkennung von Abweichungen eines Signals von einem erwarteten Wert mit der Fähigkeit den erwarteten Wert an permanente Veränderungen anzupassen. Somit eignet es sich zum Beispiel für Video-Vordergrund/Hintergrund Analysen, da unerwartete Veränderungen als Vordergrund erkannt werden, aber erwartete Veränderungen, wie das Bewegen eines Baumes im Wind, als Hintergrund. Außerdem werden permanente Veränderungen, wie ein neu abgestelltes Fahrzeug, zunächst als Vordergrund markiert, jedoch über einen kurzen Zeitraum Teil des Hintergrundes. Diese Methode wird im Rahmen der AVC-Matrix Methode sowohl für Video- als auch für Audioanalyse verwendet.

Das Signal wird mit einer Kombination von R Gaußverteilungen modelliert. Dadurch ergibt sich die Wahrscheinlichkeit P den Zustand z^t zum Zeitpunkt t zu beobachten wie folgt:

$$P(z^{(t)}) = \sum_{r=1}^{R} w_r^{(t)} * \mathcal{N}(z^{(t)} | \mu_r^{(t)}, \sigma_r^{(t)}) \tag{1}$$

Wobei $\mathcal{N}$ eine Gaußverteilungen mit Gewichtung $w_r^{(t)}$, Durchschnitt $\mu_r^{(t)}$ und Standardabweichung $\sigma_r^{(t)}$ ist. All diese Variablen verändern sich im Laufe der Zeit und sind somit von t abhängig. Die Summe der Gewichte alle Gaußverteilungen muss immer eins ergeben.

Die Gaußverteilungen werden zunächst absteigend nach Gewichtung w sortiert. Die ersten Gaußverteilungen sind die "erwarteten" Werte beziehungsweise "der Hintergrund". Wird ein neuer Zustand z gemessen, wird er nacheinander gegen jede Gaußverteilung G_1 bis G_R geprüft. Ist dieser Wert nicht mehr als $2.5\sigma_i$ vom Durchschnitt μ_i entfernt, so gilt die Gaußverteilung G_i als Treffer. Anschließend werden die Paramter der Gaußverteilung G_i wie folgt geändert:

1. Ist G_i ein Treffer
   - Erhöhe die Gewichtung w_i (siehe Gleichung 2
   - Nähere den Durchschnitt μ_i an z an (siehe Gleichung **??**
   - Verringere die Standardabweichung σ_i (siehe Gleichung **??**
   - Summiere alle Gewichte der bereits überprüften Verteilungen $G_1, ..., G_i$. Ist die Summe größer als der festgelegte Schwellwert T, markiere z als Vordergrund.
2. Ist G_i kein Treffer
   - Verringere die Gewichtung w (siehe Gleichung 2
3. Ist keine Gaußverteilung ein Treffer

- Markiere z als Vordergrund
- Entferne die Letzte Gaußverteilung in der Liste
- Füge eine neue Gaußverteilung hinzu mit $\mu = z$, einer hohen Varianz σ und einer geringen Gewichtung w

Die Veränderung der Gewichtung w erfolgt mit folgender Gleichung:

$$w_r^{(t)} = (1 - \alpha) * w_r^{(t-1)} + \alpha * M^{(t)} \tag{2}$$

wobei $M^{(t)} = 1$ wenn die Gaußverteilung ein Treffer ist, andernfalls gilt $M^{(t)} = 0$. Anschließend müssen alle Gewichte normalisiert werden, dass sie in Summe eins ergeben. Je größer die Lernrate α ist, desto schneller passt sich das Modell an Veränderungen an.

α ist wie T ein konfigurierbarer Wert und sollte an die Besonderheiten der Aufnahmen angepasst werden.

Die Standardabweichung und Durchschnitt für Treffer wird mit folgenden Gleichungen angepasst:

$$(sieheGleichung\mathbf{??}\mu_{r_{treffer}}^{(t)} = (1 - p)\mu_{r_{treffer}}^{(t-1)} + pz^{(t)} \tag{3}$$

$$(sieheGleichung\mathbf{??}\sigma_{r_{treffer}}^{2(t)} = (1-p)\sigma_{r_{treffer}}^{2(t-1)} + p(z^{(t)} - \mu_{r_{treffer}}^{(t)})^T(z^{(t)} - \mu_{r_{treffer}}) \tag{4}$$

mit $p = \alpha * \mathcal{N}(z^{(t)}|\mu_{r_{treffer}}^{(t)}, \sigma_{r_{treffer}}^{(t)})$, wobei α wie in Gleichung 2 die Lernrate ist.

Es ist eine sogenannte Bootstrap Phase notwendig, in der das Modell den Hintergrund "lernt". Anschließend kann es unerwartete Ereignisse erfolgreich als Vordergrund markieren. Die Dauer der Bootstrap Phase ist unter anderem von der Größe der Lernrate α abhängig.

2.3 Videoanalyse

Die Videoanalyse besteht aus zwei separaten Modulen. Das erste extrahiert Vordergrundaktivität auf Pixelebene während das zweite Modul durch ein Farben-Histogramm neu erscheinende Objekte identifiziert, siehe Abbildung 1.

Die Vordergrunderkennung verwendet das in 2.2 beschriebene TAPPMOG Verfahren, wobei jedes Pixel $z_n^{(t)}$ einzeln von einem TAPPMOG Modell beschrieben wird. Dargestellt wird ein Pixel als Vektor bestehend aus seinen RGB Werten. Somit wird jedes Pixel unabhängig von allen anderen Pixeln entweder als Vordergrund oder Hintergrund eingestuft.

In einer Videoaufnahme, in welcher ein Baum im Wind weht, wird das selbe Pixel abwechselnd grün sein, wenn die Blätter des Baumes zu sehen sind oder blau, wenn die Blätter zur Seite geweht wurden und der dahinter liegende Himmel zu sehen ist. Nach einer kurzen Trainingszeit (Bootstrapping Phase) wird das TAPPMOG Modell eine Gaußverteilung für grün und eine für blau mit hoher Gewichtung beinhalten

und weder grüne noch blaue Pixel werden als Vordergrund markiert. Ändert sich jedoch die Farbe des Pixels auf zum Beispiel Rot weil eine Person mit roter Kleidung in der Kameraaufnahme erscheint, wird entweder keine Gaußverteilung als Treffer markiert oder die Summe der Gewichte ist größer als der Schwellwert T und somit wird das Pixel als Vordergrund markiert. Bleibt die Person dort lange Zeit stehen wird sich das Modell anpassen, einen roten Wert erwarten und das Pixel nicht länger als Vordergrund markieren. Die Dauer hierfür ist von der Lernrate α abhängig, welche in dem in [Cri07] vorgestellten Verfahren für alle Pixel und auch Audio Analysen identisch ist.

Das zweite Modul verwendet die Daten der Vordergrunderkennung um das Erscheinen und Verschwinden von Objekten in den Videoaufnahmen zu erkennen. Hierfür werden alle als Vordergrund markierte Pixel für jeden Zeitschritt t in ein Video Vordergrund Histogramm (VVGH) mit J Säulen unterteilt, wobei jede Säule $v_j^{(t)}$ einen Ausschnitt des Farbspektrums (oder einer Grauskala) repräsentiert. Nun wird jede Säule mit einem eigenen TAPPMOG Modell (siehe 2.2) analysiert um unerwartete Veränderungen zu erkennen. Eine neu erscheinende Person in einer sonst nur aus Hintergrund bestehenden Szene wird als Vordergrund markiert, wodurch die Anzahl der Vordergrundpixel im VVGH unerwartet erhöht wird. Daraus lässt sich das Erscheinen eines neuen Objektes erkennen. Bewegt sich die Person einige Zeit in der Szene ist sie weiterhin teil des Vordergrundes, aber das TAPPMOG Modell der VVGH Säulen haben sich an den neuen Wert angepasst und markieren keine Vordergrundaktivität. Verlässt die Person nun die Szene sinkt die Anzahl der Vordergrundpixel. Daraus lässt sich das Verschwinden eines Vordergrund-Objektes erkennen.

Dieses Verfahren hat den Vorteil, dass es Objekte und Personen unabhängig von der Position in der Szene identifiziert. Es hat jedoch den nachteil, dass Objekte lediglich anhand der Farbe erkannt werden, wodurch zwei gleichfarbige Objekte nicht unterschieden werden können.

2.4 Audioanalyse

Die Audioanalyse basiert darauf die zeitliche Veränderung der Energiewerte einzelner Frequenzbereiche zu analysieren. Diese Energiewerte erlauben die Erkennung und Klassifizierung von akustischen Ereignissen [Pel01].

Dafür werden die Audioaufnahmen in überlappende Zeitfenster der Länge W_a unterteilt, wobei jedes Zeitfenster für den Zeitpunkt t am $t$en Videoframe endet, wie in Abbildung 1 zu sehen ist. Für jedes Zeitfenster werden mit dem Yule-Walker Autoregressionsverfahren [Mar87] die Energiewerte (in dB) $X^{(t)}(f_n)$ für $n = 1, ..., N$ berechnet, wobei f_n die Frequenz in Herz ist. Die maximale Frequenz ist $f_N = F_s/2$ wobei F_s die Samplingrate der Aufnahme ist. Die Genauigkeit der Analyse ist Abhängig von der Anzahl an Frequenzen N und der Zeitfensterlänge W_a.

Für jedes Zeitfenster wird die *Zeitfenster Energiemenge* (ZEM) berechnet, welche das Histogramm in Abbildung 1 repräsentiert. Diese Energiemenge wird in I Frequenzbereich unterteilt, wobei jede Untermenge $a_1, ..., a_I$ eine Säule des Histogramms

darstellt. Für jede Säule wird ein eigenes TAPPMOG Model trainiert um "unerwartete" Werte und damit den akustischen "Vordergrund" zu erkennen. Die ZEM ist ausreichend aussagekräftig um Audio-Vordergrund zu erkennen und Ereignisse zu klassifizieren [Row00].

2.5 Audio-Video-Kombination

Nun soll ein Zusammenhang zwischen Audio und Video Vordergrund hergestellt werden. Hierfür werden die Säulen $a_1, ..., a_I$ des Audio-Histogramms und die Säulen $v_1, ..., v_J$ des Video-Histogramms zum Zeitpunkt t kombiniert. Ein *Audio Vordergrund Muster* $A_i^{(t^A_{init}, t^A_{end})}$ in Relation zum Zeitpunkt t ist folgendermaßen definiert:

$$A_i^{(t^A_{init}, t^A_{end})} = [a_i^{(t^A_{init})}, a_i^{(t^A_{init}+1)}, ..., a_i^{(t)}, ..., a_i^{(t^A_{end})}] \tag{5}$$

wobei $\forall t \in [t^A_{init}, t^A_{end}], a_i^{(t)} \in Vordergrund$ gelten muss. Ein *Audio Vordergrund Muster* ist somit eine Reihe von zeitlich aufeinanderfolgenden Werten vom $a_i^{(t)}$, welche alle als Vordergrund eingestuft wurden. Auf diese Art können auch *Video Vordergrund Muster* $V_i^{(t^V_{init}, t^V_{end})}$ definiert werden.

Eine zeitliche Überlappung (PRI) zweier Vordergrund Muster $A_i^{(t^A_{init}, t^A_{end})}$ und $V_i^{(t^V_{init}, t^V_{end})}$ lässt sich folgendermaßen berechnen:

$$t^{AV}_{init} = max(t^A_{init}, t^V_{init}) \tag{6}$$

$$t^{AV}_{end} = min(t^A_{end}, t^V_{end}) \tag{7}$$

Wenn $t^{AV}_{init} < t^{AV}_{end}$ geil, kann die PRI beschrieben werden als:

$$PRI_{i,j}^{(t^{AV}_{init}, t^{AV}_{end})} = [t^{AV}_{init}, t^{AV}_{init}] \tag{8}$$

$PRI_{i,j}^{(t^{AV}_{init}, t^{AV}_{end})}$ ist ein Zeitintervall in welchem die Audiosäule a_i und die Videosäule v_j gleichzeitig als Vordergrund markiert sind. Die Gewichtung der AV Kombination ist definiert als

$$w^{(t)}_{AV}(i,j) = \frac{w^{(t)}_{A,i,q} + w^{(t)}_{V,j,u}}{2} \tag{9}$$

für $t \in PRI_{i,j}^{(t^{AV}_{init}, t^{AV}_{end})}$ (ansonsten gilt $w^{(t)}_{AV}(i,j) = 0$). $w^{(t)}_{V,j,u}$ beziehungsweise $w^{(t)}_{A,i,q}$ sind die Gewichte der Gaußverteilungen für die Videosäule $v_j(t)$ beziehungsweise Audiosäule $a_i(t)$ in den jeweiligen TAPPMOG Modellen, wie in Abbildung 2 zu sehen ist.

Aus diesen Daten wird die AVC Matrix erstellt. Diese $I * J$ große Matrix beschreibt den zeitlichen Verlauf vom Zeitpunkt 0 bis t folgendermaßen:

$$AVC^{(t)}(i,j) = \sum_{t'=0}^{t} w^{(t')}_{AV}(i,j) \tag{10}$$

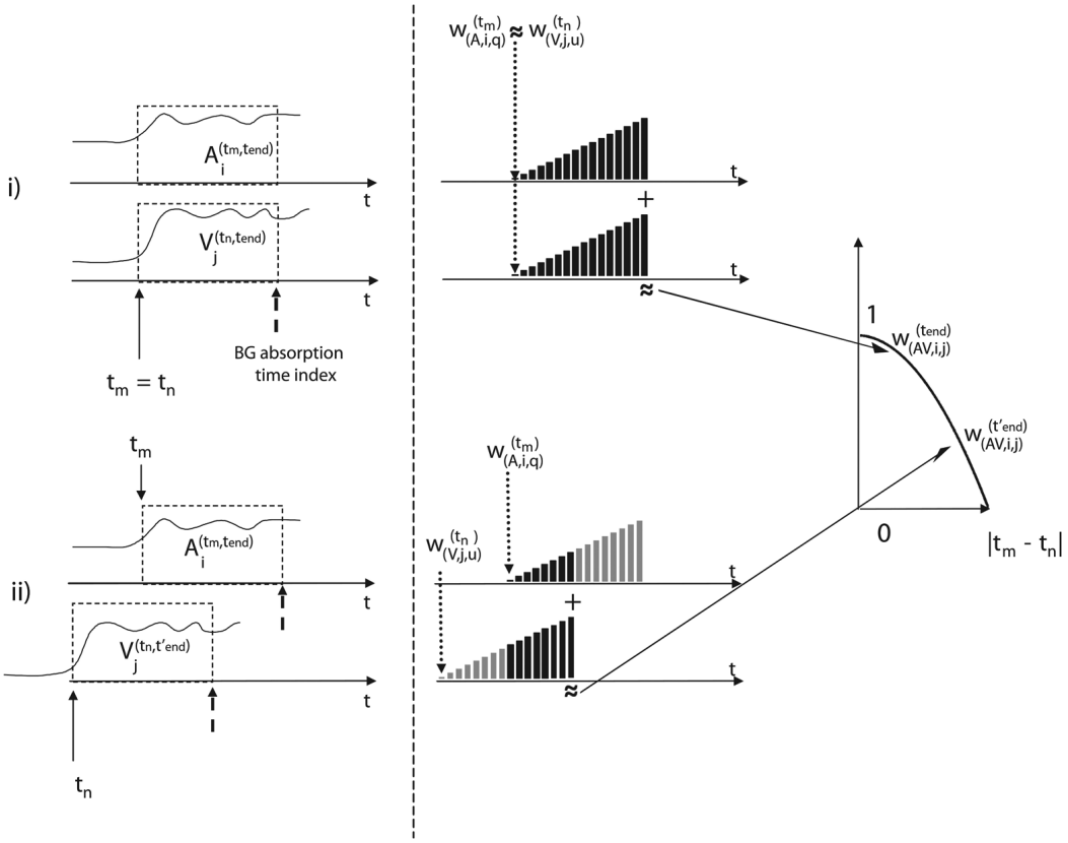

Abbildung 2: Darstellung des Audio-Video-Kombinationsverfahren. Zeile (i) zeigt zwei stark synchrone Ereignisse während (ii) zwei zeitlich verschobene Ereignisse zeigt. In der Linken Spalte sind die Vordergrund Muster für Audio und Video zu sehen. Die Markierung l für den "BG absorption time index" zeigt den Zeitpunkt ab welchem sich das System an das Vordergrund Ereignis angepasst hat und es als Hintergrund kategorisiert. Rechts der gestrichelten Linie sind die Gewichte der Gaußverteilungen der Vordergrundmuster. Das Diagramm rechts zeigt die Kombinationsgewichtung: Je synchroner die Ereignisse sind, desto höher ist ihre Gewichtung

Der Eintrag $[i, j]$ in der AVC Matrix ist die Summe aller bisherigen Audio-Video-Kombinationsgewichtungen für die Audiosäule a_i und die Videosäule v_j. Somit ist die Matrix zum Zeitpunkt $t = 0$ leer und wird mit jedem neuen Zeitschritt aktualisiert.

2.6 Event Erkennung

Ein *audiovisuelles Event* (AVE) ist im Sinne der AVC Matrix Analyse definiert als ein Zeitintervall, in dem sowohl Audio als auch Video Vordergrund festgestellt wurde. Eine Kombination der Audiosäule a_i und der Videosäule v_j existiert dann, wenn die AVC Matrix an den Koordinaten $[i, j]$ einen nicht-null Wert enthält. Ein AVE existiert dann im Zeitintervall $[t^{AV}_{init}, t^{AV}_{init}]$, wenn folgendes gilt:

$$AVC^{(t^{AV}_{init}-2)} - AVC^{(t^{AV}_{init}-1)} = 0 \tag{11}$$

$$\forall t \in [t^{AV}_{init}, t^{AV}_{init}], AVC^{(t)} - AVC^{(t+1)} \neq 0 \tag{12}$$

$$AVC^{(t^{AV}_{end}+1)} - AVC^{(t^{AV}_{end})} = 0 \tag{13}$$

Gleichung 11 besagt, dass vor Beginn des Events die AVC Matrix unverändert sein muss (also darf kein Audio- und Video-Vordergrund synchron existieren). Die Gleichung 12 besagt, dass während des Zeitintervalls $[t^{AV}_{init}, t^{AV}_{init}]$ durchgehend synchron Audio- und Video-Vordergrund existieren muss und somit bei jedem Zeitschritt mindestens ein Wert der AVC Matrix erhöht wird. Die Gleichung 13 besagt, dass nach Ende des Events keine weitere Synchronität zwischen Audio- und Videovordergrund bestehen darf.

Nun wird für ein AVE k eine *Audio Video Beschreibung* (AVB) erstellt. Diese wird aus der AVC Matrix extrahiert

$$AVB(AVE_k) = AVC^{(t^{AV}_{end}(k))} - AVC^{(t^{AV}_{init}(k)-1)} \tag{14}$$

Die AVB enthält die AV Informationen, welche während des Events k gesammelt wurden. Diese Information wird vektorisiert und kann zur Identifizierung von audiovisuellen Events verwendet werden, wie in 2.7 gezeigt wird.

2.7 Experimentelle Ergebnisse

Um die Aussagekräftigkeit der in dieser Methode gewählten AV Eigenschaften zu testen, wurde sie in [CBM] anhand von echten Kameraaufnahmen getestet. Hierbei ist entscheidend, dass die Methode die Events nicht zu allgemein oder zu spezifisch kategorisiert und somit Ergebnisse ähnlich wie ein menschlicher Analyst liefert.

Die Testdaten stammen aus einem Büro und enthalten einzelne Events. Die Events bestehen aus alltäglichen Büroaktivitäten, wie zum Beispiel das Büro betreten oder verlassen, einen Telefonanruf entgegennehmen oder das Bürolicht ein- oder ausschalten (siehe Abbildung 3). Die einzelnen Events ereignen sich in einem Abstand von 0.5 bis 10 Sekunden. Die Events überlappen sich nicht, somit ist ein Event immer vollständig

Abbildung 3: Aufnahmen aus den Testdaten

abgeschlossen, bevor ein neues beginnt. Es wurden zwei Datensets mit jeweils einer Länger von über zwei Stunden aufgenommen, wobei die aufgenommene Person für das zweite Datenset mehrfach die Kleidung wechselte um die Varianz zu erhöhen.

Die Videos wurden mit einer 320x240 CCD Kamera mit 20 Bildern pro Sekunde aufgenommen. Die Audioaufnahmen wurde mit 22050Hz erstellt und in Zeitfenster der Länge $W - a = 1s$ mit einer zeitlichen Überlappung von 70% unterteilt. Für die Anzahl I und J der Audio- und Videosäulen in den Histogrammen wurde $I = J = 8$ gewählt. Versuche zeigten, dass diese Parameter ab einer Größe von 32 zu Performance Problemen führen. Der Wert 8 erlaubt es alle Berechnungen nahezu in Echtzeit durchzuführen mit einem Pentium III 500MHz Prozessor in MATLAB und gleichzeitig eine ausreichende Genauigkeit zu liefern.

Die Audioaufnahmen wurden in $I = 8$ gleichgroße Abschnitte zwischen $[0, 22050/4]hz$ unterteilt. Für jeden Abschnitt (ein Abschnitt entspricht einer Säule des Histogramms) wurde eine Kombination von drei Gaußverteilungen gewählt (siehe 2.2). Diese Anzahl ist für den einfachen Aufbau der Szenen in den Testdaten ausreichend. Für den Schwellwert T wurde 0.8 gewählt und für die Lernrate α 0.001. Die initiale Gewichtung der Gaußverteilungen für alle TAPPMOG Modelle ist $w_{init} = 0.001$. Diese Werte wurden durch einige initiale Testdurchläufe als akzeptable befunden.

Die Videoaufnahmen werden um den Faktor vier komprimiert, bevor die Vordergrunderkennung auf Pixelebene jeweils mit einem TAPPMOG Modell mit drei Gaußverteilungen durchgeführt wird. Der Vordergrund wird anschließend in ein Histogramm mit $J = 8$ Säulen umgewandelt. Jede Säule stellt einen gleichgroßen Teil des Grauspektrums $[0, 255]$ dar. Für jede Säulen wird ein TAPPMOG Modell mit drei Gaußverteilungen verwendet. Für die Vordergrunderkennung und Histogrammanalyse wird

der gleiche Schwellwert T verwendet, jedoch eine andere initiale Standardabweichung σ_{init} für neu initialisierte Gaußverteilungen (siehe 2.2 Fall 3. "Keine Gaußverteilung ist ein Treffer"), da sich die Zahlenintervalle stark unterscheiden.

Die Größe der Standardabweichung stellte sich als entscheidend für die AVC Methode heraus. Bei einem zu geringen Wert deckten mehrere Gaußverteilungen gleiche Muster ab, wodurch andere Hintergrundmuster fälschlicherweise als Vordergrund markiert wurden. Bei einer zu großen Standardverteilung ist die Gefahr groß, dass eine Gaußverteilung mehrere verschiedene Muster gleichzeitig abdeckt. Nach der Analyse mehrere Konfigurationen wurden folgende Werte für die initialen Standardabweichungen verwendet:

1. $\sigma^A_{init} = 10$ für die Audioanalyse (ZEM Wert im Bereich $[0, 150]$)
2. $\sigma^P_{init} = 30$ für die Pixel-Vordergrundanalyse (Pixelsignal im Bereich $[0, 255]$)
3. $\sigma^V_{init} = 50$ für die Videohistogramm-Analyse (VVGH Werte im Bereich $[0, 4800]$

In Abbildung 4 ist der Ablauf eines AV Events AVE zu sehen. Nach dem Ende des AVEs kann die AV Beschreibung AVB als Vektor aus der Veränderung der AVC Matrix berechnet werden. Dieser AVB Vektor wird als Eingabe für Klassifizierung und Clustering verwendet.

Die kompletten vier Stunden Testdaten wurden zunächst von einem menschlichen Analysten bewertet, welcher 66 AVEs erkannte. Diese wurden kategorisiert in folgende Ereignisse:

1. *Anruf tätigen*: Eine Person bewegt sich zum Telefon, wählt eine Nummer und führt ein Gespräch
2. *Anruf erhalten*: Das Telefon klingelt, eine Person bewegt sich zum Telefon, hebt ab und führt ein Gespräch
3. *Ankunft (erster)*: Eine Person betritt das dunkle Labor, schaltet das Licht an und betritt den Raum ohne zu sprechen.
4. *Ankunft (nicht erster)*: Eine Person betritt den beleuchteten Raum und spricht
5. *Verlassen (letzter)*: Eine Person verlässt das Labor ohne zu sprechen und schaltet das Licht aus
6. *Verlassen (nicht letzter)*: Eine Person verlässt das Labor und spricht

Um die Klassifizierungsgenauigkeit der AVC Matrix Analyse zu bewerten, wurden folgende Szenarien genauer untersucht:

- Szenario A - Situation 1 und 2 unterschieden (Anruf tätigen / erhalten)
- Szenario B - Situation 3 und 4 unterscheiden (Leeres / nicht-leeres Labor betreten)

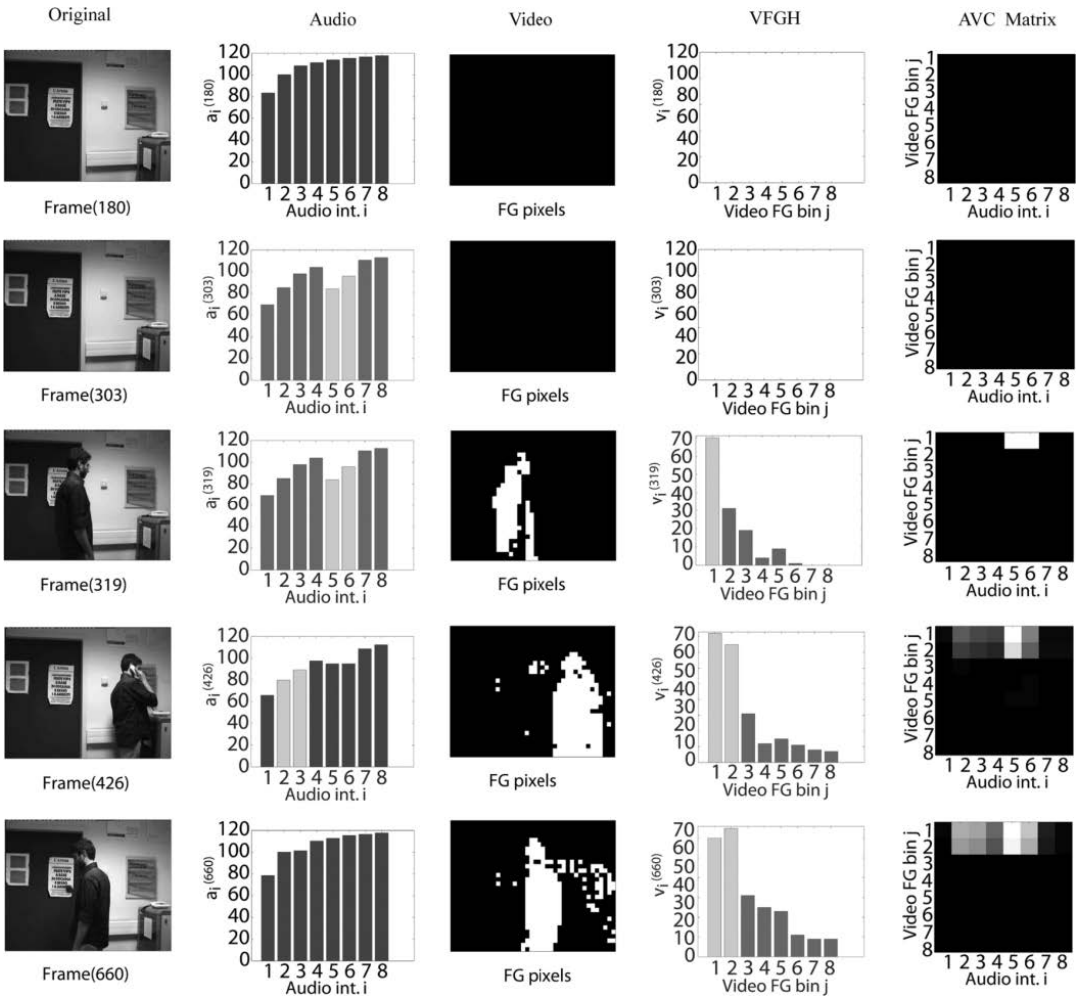

Abbildung 4: Beispiel: “Anruf erhalten” Event. Das Telefon im Bild klingelt, eine Person betritt das Bild, beantwortet den Anruf und verlässt danach das Bild. Jede der fünf Zeilen enthält einen Frame und dessen AVC Matrix Analyse:

1. Frame 180: Keine Aktivität, der Hintergrund ist etabliert
2. Frame 303: Das Telefon klingelt und verursacht Audiovordergrund (Balken 5 & 6 im Audiohistogram). Da jedoch kein Videovordergrund erkannt wurde bleibt die AVC Matrix leer.
3. Frame 319: Eine Person betritt das Bild und wird als Vordergrund erkannt. Die Kombination aus Vordergrund in der Videosäule 1 und den Audiosäulen 5 & 6 erzeugt Einträge in der AVC Matrix an Position [5,1] und [6,1]. Der Anfang es AV Event wird erkannt.
4. Frame 426: Die Person spricht am telefon und die AVC Matrix wird weiter befüllt.
5. Frame 660: Die Person hat das Gespräch beendet. Da kein Audiovordergrund mehr existiert bleibt die AVC Matrix unverändert und das Ende des AV Events wird erkannt.

Tabelle 1: Erfolgsquote beim Klassifizieren der Szenarien anhand von Audio, Video oder Audio-Video

Szenario	Audio	Video	AV Kon.	AVC Matrix
A	100,00%	86,35%	-	100,00%
B	60,87%	95,65%	-	95,65%
C	95,24%	85,71%	-	95,24%
D	62,12%	66,67%	82,28%	89,39%

- Szenario C - Situation 5 und 6 unterscheiden (Leeres / nicht-leeres Labor verlassen)
- Szenario D - Alle Situationen unterscheiden

Die Klassifizierung wurde mit dem KNN Algorithmus (Euklidische Distanz) berechnet [Dud01]. Dies ist ein simpler Klassifizerungsalgorithmus, erlaubt es jedoch die Aussagekräftigkeit von Eigenschaften zu bewerten. Die Bewertung der Ergebnisse wurde mit dem "Leave-One-Out" (LOO) Verfahren durchgeführt [Dud01]. Die Klassifizierung anhand der AVC Matrix wurde verglichen mit der Klassifizierung anhand des Audio- beziehungsweise Videovordergrund alleine. Hierfür wurde die Klassifizierung statt mit dem Vektor der AVB aus der AVC Matrix direkt mit dem Vektor aus dem Audio- beziehungsweise Video-Histogramm durchgeführt.

Die Audioanalyse ist sehr effektiv in Szenario A, da das Erhalten eines Anrufes anhand des Klingelgeräusches deutlich vom Tätigen eines Anrufes zu unterscheiden ist. Die reine Videoanalyse liefert in Szenario B und C ein zufriedenstellendes Ergebnis, da sich die Situationen durch die an- beziehungsweise ausgeschaltete Beleuchtung visuell stark unterscheiden. In jedem Fall ist die Verwendung der AVC Matrix, statt nur einer einzelnen Modalität, gleich erfolgreich oder erfolgreicher. Für Szenario D, Kategorisierung aller Situationen, ist die AVC Matrix Analyse sogar etwa 25% erfolgreicher. Ein wichtiger Vergleich ist noch mit der Leistung der AV Konkatenation ("AV Kon." Spalte in Tabelle 1). Bei dieser Methode werden die Vektoren von Audio und Video zu einem einzelnen Vektor zusammengefasst, ähnlich der in [Bar03] verwendeten Methode. Auch hier ist die AVC Matrix Analyse erfolgreicher, wodurch gezeigt wird, dass die zeitliche Synchronität der beiden Modalitäten eine nutzbringende Information ist.

Zur weiteren Bewertung der Methode wurden die Events hierarchisch, mit euklidische Distanz in Cluster gruppiert [Jai88] , wie in Abbildung 5 zu sehen ist. An der X-Achse des Diagramms ist die Nummer der Situation des Events zu sehen. Situation 2, 3 und 5 wurden sehr gut erkannt, da diese sehr eindeutige Merkmale besitzen: Ein klingelndes Telefon beziehungsweise das ein- oder ausschalten der Beleuchtung. Situation 4 und 6 sind schwerer unterscheidbar: In Situation 4 gibt es einen AV Vordergrund wenn die Tür geöffnet wird und die Person den Raum betritt. Der anschließende Audiovordergrund beinhaltet das Schließen der Tür und das Sprechen der Person während die Person sich durch die Szene bewegt. In Situation 6 hingegen ist nach dem Schließen der Tür

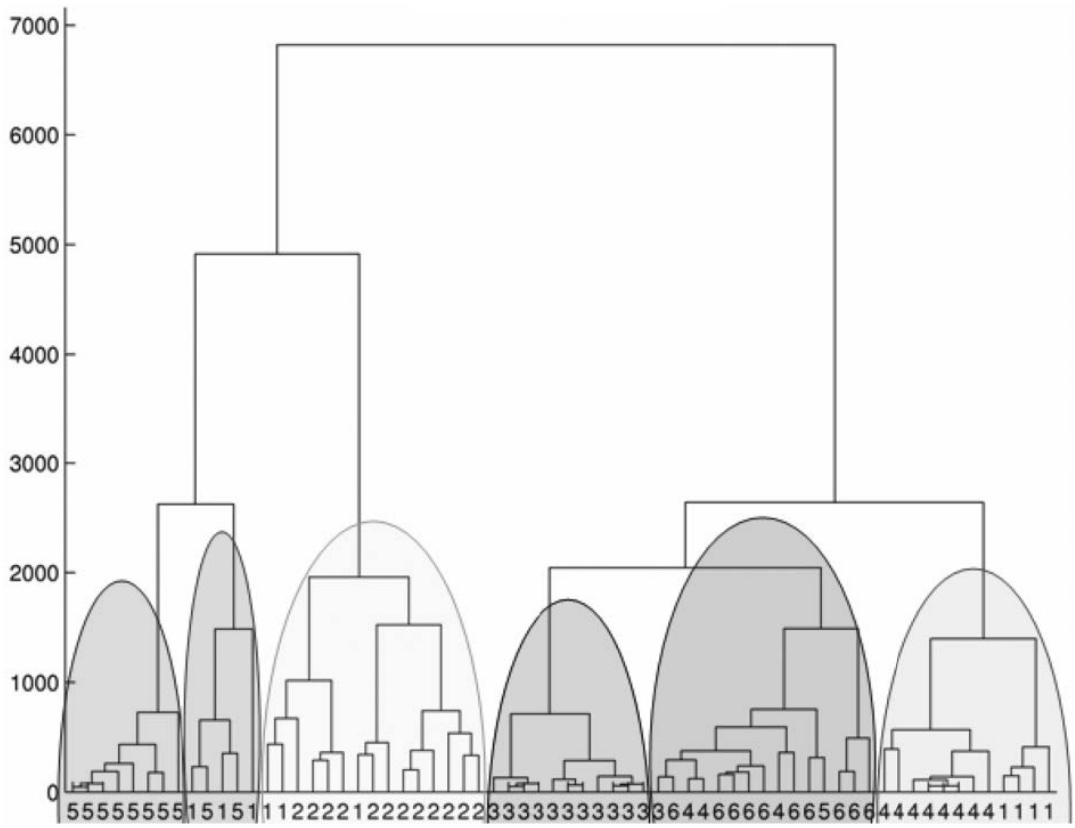

Abbildung 5: Clustering der Daten in einem Dendrogram

kein Videovordergrund mehr in der Szene vorhanden, wodurch das Schließgeräusch nicht in die AVC Matrix eingerechnet wird.

Die Clustering Genauigkeit wird aus der Anzahl der Ereignisse in falschen Clustern berechnet, wobei die Nummer des Clusters gleich der Nummer der Mehrheit seiner Einträge ist. Somit beträgt die Genauigkeit 75,76%. Im Vergleich zum reinen Audio Clustering mit 53,03%, reinem Video Clustering mit 60,61% und dem Clustering anhand der einfachen Videokonkatenation mit 64,44% liefert die AVC Matrix Methode ein deutlich genaueres Ergebnis und zeigt, dass die Verbindung der Audio- und Video-Modalitäten und deren zeitliche Synchronität einen höheren Informationsgehalt liefern, als bisherige Methoden.

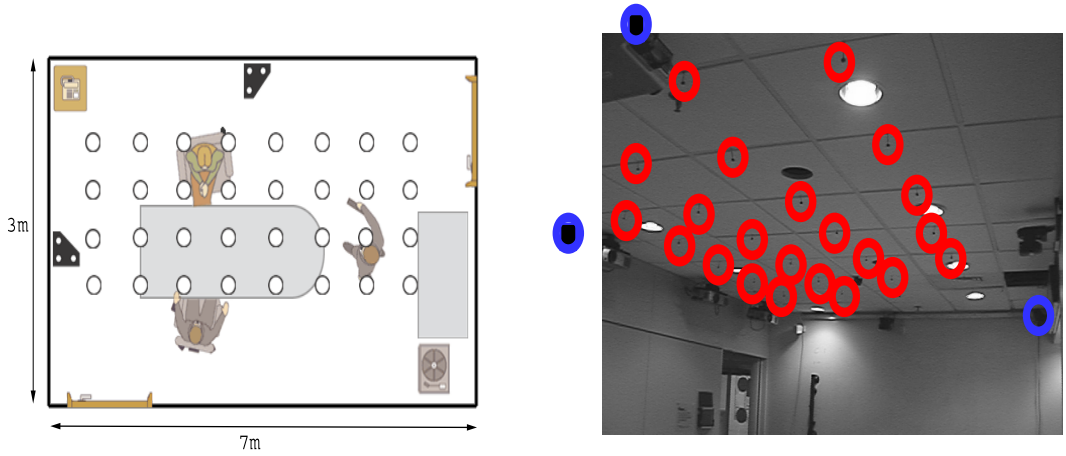

Abbildung 6: Links eine schematische Zeichnung, in der die Kreise Mikrofone darstellen und die schwarzen Dreiecke Kameras. Rechts ist ein Foto der Testumgebung in welchem Kameras und Mikrofone hervorgehoben wurden. Quelle [Wil01b]

3 Alternative Methoden

3.1 Mikrofonreihen

Ein System bestehend aus mehreren Mikrofonen und Kameras ermöglicht eine sehr detaillierte Eventerkennung und räumlich Bewegungsverfolgung von Objekten und Personen im überwachten Bereich. Das in [Che02] beschriebene System überwacht einen Konferenzsaal mit 32 ungerichteten Mikrofonen, welche an der Decke des Saales befestigt sind, und zwei Videokameras ähnlich zu der in Abbildung 6 dargestellten Umgebung.

Die Videoaufnahmen werden mit einer Foreground-Background-Analysis aufbereitet um Veränderungen in den Aufnahmen kenntlich zu machen. Diese Veränderungen im Verhältnis zur Zeit stellen die Bewegungen der Personen in den Videoaufnahmen dar [Dar01]. Da die Anwendung Personen in einem Konferenzsaal beobachtet, wird die Position lediglich auf eine zweidimensionalen Fläche projiziert. In Abbildung 7 ist eine Videoanalyse zu sehen. Die X und Y Achsen geben die Position der Bewegung im Raum an während die Z Achse der zeitliche Verlauf ist. In der Abbildung ist eine Momentaufnahme für t=29 zu sehen. Diese zeigt alle Regionen, in denen zu diesem Zeitpunkt t Bewegungen festgestellt wurden. Insgesamt zeigt die Grafik einen Bewegungsverlauf von zwei Personen, welche sich der Raummitte nähern und einem Objekt in der Raummitte, welches ungefähr ab dem Zeitpunkt t=25 bewegt wird. Nach dem Zeitpunkt t=40 bewegen sich die zwei Personen wieder aus der Mitte des Raumes hinaus.

Mit den Werten der Mikrofone wird die Lautstärke im Vergleich zu Rauschgeräuschen (SNR, Signal-Noise-Ratio) auf dem Raum abgebildet. Somit wird ein Graph wie in Abbildung 8 erzeugt. Da dieses System eine Gruppe von Menschen beobachtet, arbeitet

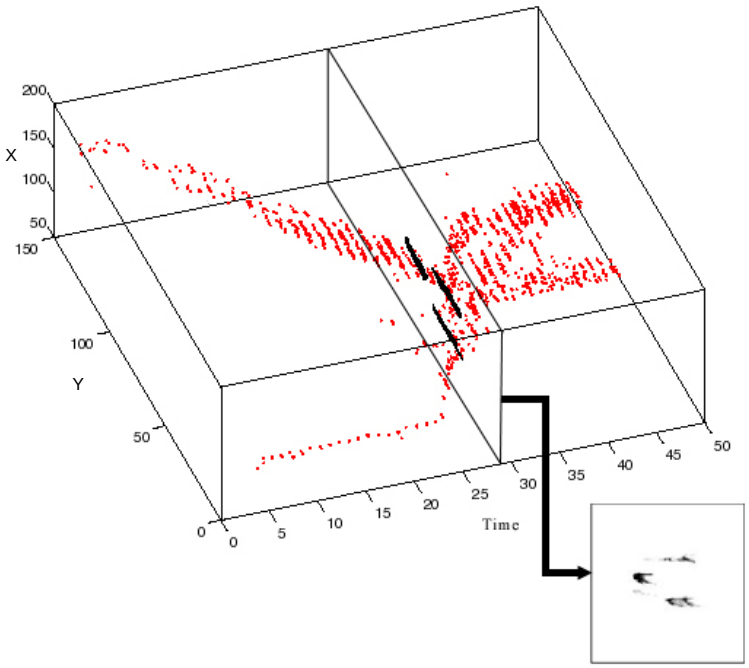

Abbildung 7: Video Analyse von Bewegungen im Verlauf der Zeit

es unter der Annahme, dass Bewegung und Geräusche korrelieren. Deshalb werden die Positionen, an denen es visuelle Bewegungen gab, als Ausgangspunkt für eine Suche nach einem lokalen Maximum im Audiographen verwendet [Che]. Somit kann die genaue Position der Geräuschquelle lokalisiert werden. Mithilfe der Position können die Zeitverzögerungen der Audioaufnahmen der unterschiedlichen Mikrofone berechnet und somit das Zielgeräusch von Hintergrundgeräuschen getrennt werden [Wil01a].

Die Audio- und Videodaten werden zu einem beobachteten Zustand z kombiniert. Dann wird berechnet wie wahrscheinlich diese Beobachtung z für eine hypothetische Konfiguration O_t

$$O_t = (o_1, ..., o_n) \tag{15}$$

aus n Objekten zum Zeitpunkt t ist, wobei $O_i = [x,\ y,\ h,\ f]$ der Zustandsvektor eines Objektes ist. Hierbei werden Personen als Zylinder der Höhe h mit festgelegten Radio dargestellt, dessen Position auf dem Boden durch $[x,\ y]$ dargestellt wird. f ist die Frequenz ihrer Stimme.

Für ein einzelnes Objekt gibt $L(z|O_i)$ an, wie wahrscheinlich die Hypothese eines einzelnen Objektes von der Beobachtung unterstützt wird. Dies wird berechnet aus

$$L(z|o_i) = L(z_a|o_i) * L(z_v|o_i) \tag{16}$$

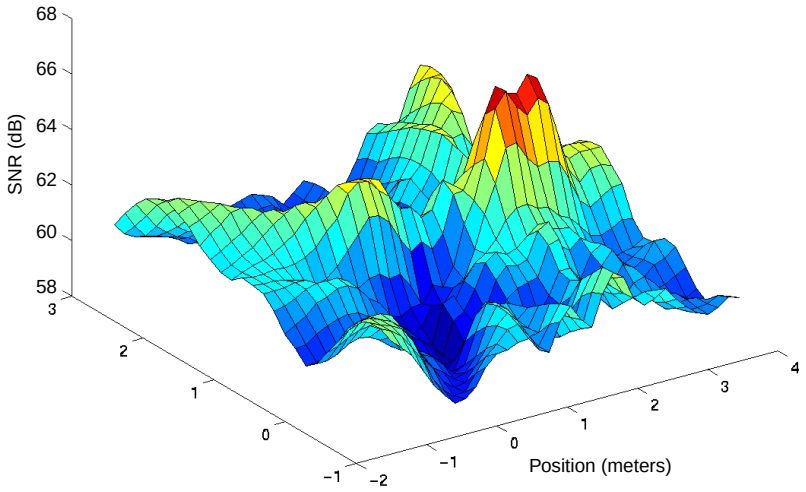

Abbildung 8: Räumliche Audio Analyse durch mehrere Mikrofone

wobei $L(z_a|o_i)$ und $L(z_v|o_i)$ angeben, wie sehr die Audio- beziehungsweise Videodaten die Hypothese o_i unterstützen.

Die Wahrscheinlichkeit $L_t(z|O)$, dass der beobachtete Zustand für eine Hypothese O bestehend aus m Objekten auftritt lässt sich mit dem gauschen Fehlintegral aus den Einzelwahrscheinlichkeiten der Objekte berechnen

$$L_t(z|O) = \phi L(z|o_1), ..., L(z|o_m) \tag{17}$$

Mikrofonreihen erlauben es selbst sich bewegende Audio-Visuelle Ereignisse mit einer Genauigkeit von bis zu 10 Zentimeter zu erkennen und verfolgen. Somit können auch Gespräche zwischen zwei nahestehenden Personen korrekt erkannt werden. Auch können zwei gleichzeitige, unabhängige AV Events separat erkannt werden. Jedoch benötigt diese Methode eine große Menge an Mikrofonen und Rechenleistung und ist daher nur in sehr kontrollierten Umgebungen einsetzbar [Che02].

3.2 Dualmikrofon mit HHM

3.3 Cononical correlation analysis

3.4 Maximization of mutual information

3.5 Computational Auditory Scene Analysis CASA

3.6 Computational Auditory Scene Recognition CASR

4 Bewertung

Die AVC Matrix Methode ist eine neue Methode um audiovisuelle Eventerkennung für Videoaufnahmen mit einer Kamera und einem Mikrofon zu automatisieren. Die Video- und Audiosignale werden separat mit adaptiven Modellen verarbeitet, um Vordergrund und Hintergrund zu unterscheiden. Anschließend werden anhand ihrer zeitlichen Synchronität Audio-Video-Events erkannt. Dies wird mit einer Audio-Video-Concurrency (AVC) Matrix berechnet. Die Methode lieferte in den durchgeführten Experimente bessere Ergebnisse, als die Verwendung reiner Audio- oder Videodaten. Auch liefert sie bessere Ergebnisse als die Verwendung von Audio-Video-Daten welche ohne die zeitliche Synchronität kombiniert wurden.

Die Methode verwendet simple Algorithmen für die Videoerkennung, wodurch zum Beispiel mehrere gleichfarbige Objekte nicht unterschieden werden können. Die AVC Matrix Methode kann jedoch auch mit weiterentwickelten Videoanalyse Methoden kombiniert werden. Außerdem ist diese Methode nicht in der Lage mehrere gleichzeitig erscheinende Events korrekt zu trennen.

Literatur

[Bar03] M. Barnard, J.-M. Odobez und S. Bengio. Multi-Modal Audio-Visual Event Recognition for Football Analysis, 2003.

[CBM] [Cri07], Section IV-A.

[Che] [Che02], Section 4.

[Che02] N. Checka und K. Wilson. Person Tracking Using Audio-Video Sensor Fusion, 2002.

[Cri07] M. Cristani, M. Bicego und V. Murino. Audio-Visual Event Recognition in Surveillance Video Sequences, 2007.

[Dar01] T. Darrell, D. Demirdjian, N.Checka und P. Felzenszwalb. Plan-view trajectory estimation with dense stereo background models, 2001.

[Dud01] R. Duda, P. Hart und D. Stork. *Pattern Classification*. Wiley, 2001.

[Jai88] A. Jain und R. Dubes. Algorithms for Clustering Data, 1988.

[Mar87] S. Marple. *Digital Spectral Analysis*. Prentice-Hall, 1987.

[Nie02] E. Niebur, S. Hsiao und K. Johnson. Synchrony: A neuronal mechanism for attentional selection, 2002.

[Pel01] V. Peltonen. Computational auditory scene recognition, 2001.

[Row00] S. T. Roweis. One Microphone Source Separation, 2000.

[Sta98] C. Stauffer und E. Grimson. Adaptive background mixture models for real-time tracking, 1998.

[Wil01a] K. Wilson, N. Checka, D. Demirdjian und T. Darrell. Audio-video array source localization for perceptual user interfaces, 2001.

[Wil01b] K. Wilson, N. Checka, D. Demirdjian und T. Darrell. Audio-Video Array Source Separation for Perceptual User Interfaces, 2001.