Visuelle Personenverfolgung mit Partikelfiltern

Studienarbeit

Institut für Logik, Komplexität und Deduktionssysteme
Prof. A. Waibel

Fakultät für Informatik
Universität Karlsruhe (TH)

von

Christian A. Wojek

30. April 2004

Betreuer:

Prof. Dr. Alex Waibel
Dr.-Ing. Rainer Stiefelhagen
Dipl.-Inform. Kai Nickel

Hiermit erkläre ich, die vorliegende Arbeit selbständig erstellt und keine anderen als die angegebenen Quellen verwendet zu haben.

Karlsruhe, 30. April 2004 .....

Inhaltsverzeichnis

1 Bedeutung visueller Personenverfolgung

1.1 Anwendungen in intelligenten Räumen

In den letzten Jahren wurde sehr viel Forschungsarbeit in die Entwicklung sogenannter intelligenter Räume investiert. Dabei stellt das sichtgestützte Verfolgen von Personen eine wichtige Grundlage dar, um weitergehende Dienste wie zum Beispiel das Steuern der Raumausstattung durch Zeigegesten anzubieten.
Denn nur wenn die Kontextinformation vorliegt, wer der Sprecher ist und wo er sich befindet, können entsprechende Befehle richtig interpretiert und die dazu passenden Aktionen eingeleitet werden. Ist dies nicht der Fall, könnten missverstandene Bewegungen von Zuhörern im Publikum dazu führen, dass das genaue Gegenteil, eine Störung des Vortragenden erreicht wird.
Des Weiteren kann die Information über den Standpunkt einer Person dazu verwendet werden, um ihre Aktivitäten zu klassifizieren. So ist es wesentlich wahrscheinlicher, dass ein an der Tafel stehender Sprecher Notizen an dieser macht verglichen mit einem Sprecher, der sich in der Mitte des Raums befindet.
Eine weitere Anwendung besteht für die Spracherkennung beim Beamforming [18]. Dabei wird die Kenntnis der Sprecherposition dahingehend ausgenutzt, dass man Geräusche, die nicht aus dessen Richtung ans Mikrophon gelangen, dämpft und dadurch die Sprecherstimme besser vom Hintergrundrauschen separiert werden kann.
In zukünftigen Anwendungen soll das entwickelte System weiter ausgebaut werden, wozu neben der visuellen Modalität außerdem noch Lokalisationsinformationen, die durch den Einsatz eines Microphone Arrays gewonnen werden, ausgenutzt werden sollen. Durch eine Hinzunahme der audiogestützen Messhypothesen sollte es möglich sein, die Trackinggenauigkeit und -zuverlässigkeit weiter zu verbessern.
Die genannten Anwendungen könnten zwar auch dadurch realisiert werden, dass die zu verfolgende Person mit zusätzlichen Kleinstcomputern ausgestattet wird, wie dies im Bereich des Ubiquitous Computing geschieht, allerdings wird auch hierfür Infrastruktur benötigt und das System könnte nicht sofort benutzt werden.
Diese Arbeit wurde im Umfeld des CHIL (Computers in the Human Interaction Loop) EU-Projekts durchgeführt, welches die Koordination verschiedener Projekte im Bereich intelligenter Räume zum Ziel hat (http://www.chil.server.de). Dabei

sind mehrere Partner beteiligt, deren Ziel es ist, nicht nur Mensch-Maschine Interaktionssysteme zu entwickeln, sondern Systeme, die in der Lage sind, zwischenmenschliche Aktivitäten im Raum zu beobachten, zu interpretieren und daraufhin implizite Dienste bereit stellen. Dies soll durch die Ausnutzung möglichst vieler Kontextinformationen erreicht werden, so dass sich die Frustration von Benutzern gegenüber Computern verringert.

1.2 Stand der Forschung

Verschiedenste Forschungsgruppen haben sich bereits mit dem Problem der Personenverfolgung beschäftigt. Unter anderem ist hier der Ansatz von Haritiaoglu [9] W^4 zu erwähnen, das in der Lage ist, mehrere Personen in Echtzeit zu verfolgen genauso wie Darells [4], dessen Stereokamerasystem Gesichter verfolgt. Daneben existiert das wohlbekannte pfinder System von Wren [19], das Körperteile einzelner Benutzer verfolgen kann.
Erheblich weniger Arbeiten wurden zu Systemen veröffentlicht, die wie diese Arbeit mehr als eine Kamera verwenden, was vermutlich auf die Tatsache zurückzuführen ist, dass das Korrespondenzproblem noch immer sehr schwer und nur unter massivem Einsatz von Rechnerkapazität zu lösen ist. Darunter versteht man die richtige Zuordnung segmentierter Bereiche aus verschiedenen Kameraansichten, die zur gleichen Person gehören. Problematisch wird dies insbesondere dann, wenn bei der Segmentierung mehr als ein Bereich pro Person errechnet wird, da die Laufzeit meist exponentiell mit der Zahl der segmentierten Regionen ansteigt. Die meisten Autoren verwenden kalibrierte Kameras, wie auch Cai and Aggarwal [2], die Punkte auf der Mittelachse von Personen verfolgen und den Suchraum durch die Einführung von Epipolarlinienbedingungen einschränken. Auch Mikic et al. [14] verwenden kalibrierte Kameras und geometrische Nebenbedingungen, um Menschen in Echtzeit zu verfolgen. Krumm et al. [13] hingegen verwenden farbbasierte Einschränkungen, um Leute in einem Wohnzimmer zu verfolgen.
Weitere Forschungsarbeit zur Verfolgung von Personen mit kombinierten Audio- und Videohypothesen und der Hilfe von Partikelfiltern wurde von Zotkin et al. [21] durchgeführt. Ihr System ist in der Lage, sowohl Menschen in einem Konferenzszenario wie auch Fledermäuse zu verfolgen, wobei das System außerdem vermag, sich selbst zu kalibrieren. Dies wird durch die Hinzunahme von Dreh- und Kippwinkel einer Kamera in den Statusvektor des Partikelfilters erreicht.
Checkas System [3] ist in der Lage, unter Ausnutzung von Audiomessungen den aktiven Sprecher zu verfolgen, wobei der Statusvektor neben der Zahl der verfolgten Personen und den Ortskoordinaten aller Personen außerdem einen Eintrag enthält ob eine Person spricht oder schweigt. Dabei kann das System eine sich verändernde Anzahl von Leuten verfolgen und nutzt sowohl die Audio- wie auch die Videomodalität aus, um ein möglichst genaues Ergebnis zu erreichen.
Im Vergleich zu vorangegangenen Arbeiten werden für das hier vorgestellte System

nicht synthetische Daten zur Evaluation verwendet, sondern Daten, die während wöchentlichen Seminarvorträgen unter realistischen Bedingungen am Interactive Systems Lab aufgenommen wurden. Hierbei gestaltet sich die Verfolgung des Sprechers insbesondere deshalb schwierig, weil sich außerdem noch ungefähr zehn weitere Zuhörer im Raum befinden. Um Referenzpositionen des Sprecherstandorts für die Evaluation zu erhalten wurden diese manuell markiert.

2 Teilkomponenten des Trackingsystems

In diesem Kapitel soll der Aufbau des implementierten Personenverfolgungssystems genauer dargestellt werden. Dabei wird zunächst auf die Ausstattung des *Smartrooms* und die verwendeten Kameras eingegangen. Weiterhin werden die Algorithmen zur Segmentierung, zur Triangulation und zur Spurglättung der erhaltenen Daten näher vorgestellt.

2.1 Einrichtung des Smartroom

Mit *Smartroom* wird am Interactive Systems Lab ein Raum bezeichnet, der sich zukünftig von anderen Besprechungszimmern dadurch unterscheiden soll, dass er intelligent auf die Bedürfnisse des Sprechers und der Zuhörer eingehen kann. Dabei soll der Benutzer keine zusätzlichen Hilfsmittel benötigen, sondern jegliche Aktion soll bereits ad-hoc durch die installierten Systeme ermöglicht werden. Dies kann zum Beispiel das Steuern der Lichtverhältnisse durch Zeigegesten sein [15] oder das automatische Protokollieren eines Treffens.

Das Zimmer ist ein etwa 6x7 Meter großer Raum, der neben den vier Firewire-Kameras mit fester Brennweite in den Ecken außerdem ein Microphone Array enthält. Dieses soll in zukünftigen Systemen die Fusion von Audio- und Videodaten ermöglichen, so dass durch die Hinzunahme einer audiogestützten Aufenthaltshypothese die Trackingergebnisse weiter verbessert werden können.
Die Kameras befinden sich in etwa zwei Meter Höhe und sind mit Hilfe der *Camera calibration toolbox for matlab* kalibriert [1] worden. Dieser liegt das Kalibrierungsverfahren von Zhang [20] zugrunde. Modelliert werden neben den Weltkoordinaten der Kamera zudem die intrinsischen Parameter, die die Verzerrung durch die Linsen mathematisch beschreiben. Details hierzu finden sich neben den oben bereits erwähnten Artikeln außerdem bei Focken [6].
Für zukünftige Anwendungen wird der *Smartroom* zusätzlich mit vier Stereokameras in den Ecken und einer dreh- und schwenkbaren Zoomkamera an der Decke ausgestattet werden. Dies wird zusätzlich zur Lokalisation von Personen im Raum die Möglichkeit eröffnen auch deren Gesten zu erkennen und so die Interaktivitätsmöglichkeiten weiter verbessern. Des Weiteren soll eine Middleware

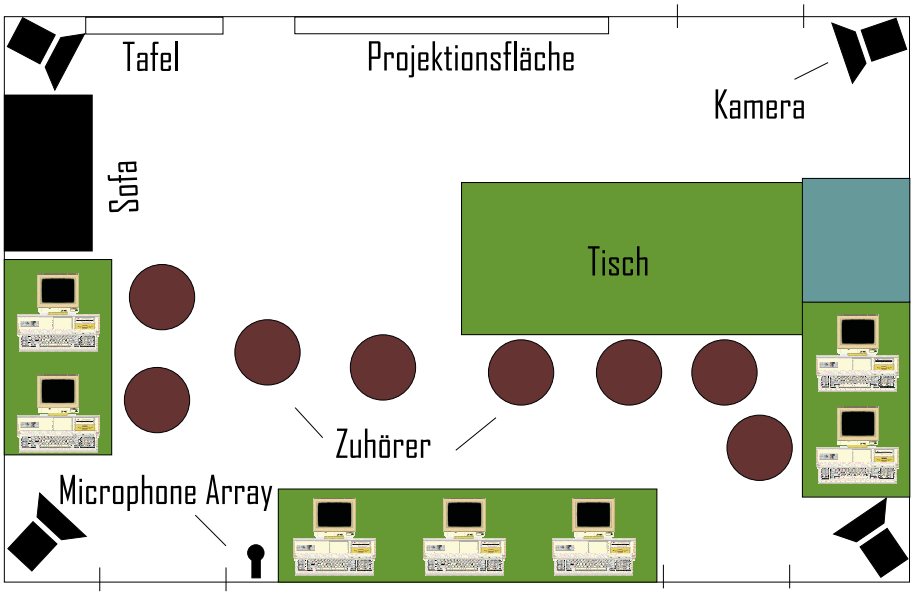

Abbildung 2.1: Schematischer Aufbau des *Smartroom* während eines Seminars

verwendet werden, die das dynamische Zu- und Abschalten von Datenströmen und modularen Diensten ermöglicht. Die in dieser Arbeit verwendeten Bilddaten wurden in verschiedenen Seminarvorträgen am Institut für Logik, Komplexität und Deduktionssysteme mit einer Auflösung von 640x480 Pixeln aufgenommen, wobei die Farbinformation mittels RGB kodiert wurde.

2.2 Segmentierung

Für die Segmentierung der Kamerabilder wird ein sogenanntes adaptives Hintergrundmodell verwendet. Details hierzu sind dem Artikel von Stauffer [17] sowie der Studien- und Diplomarbeit von Focken [6, 5] zu entnehmen.
Die wesentliche Eigenschaft besteht darin, dass für alle vier Kameras für alle Pixel die auftretenden Farbwerte als eine Mixtur von Normalverteilungen modelliert werden. Dabei werden alternativ jeweils der Rot-, Grün- und Blau- Kanal oder in der Implementierung von Focken der Helligkeitswert Y und die beiden Kanäle des rg-Farbraums getrennt voneinander betrachtet. Die Farbinformation des r und g Kanals errechnet sich dabei wie folgt:

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B}$$

Da es keine Lichtquellen mit sich zyklisch ändernden Farben im Hintergrund gibt, ist die Modellierung durch eine einzige Normalverteilung für die jeweiligen Farbkanäle mit Mittelwert μ und Standardabweichung σ ausreichend.

Man macht zudem folgende Annahmen:

- Der Hintergrund wird die meiste Zeit beobachtet, so dass μ für alle Kanäle nahe am Farbwert des tatsächlichen Hintergrunds liegt.
- Der zu verfolgende Sprecher unterscheidet sich deutlich vom Hintergrund im beobachteten Farbwert und bewegt sich in regelmäßigen Abständen, so dass er nicht zum Hintergrund adaptiert wird.
- Die beobachteten Farbwerte der verschiedenen Kanäle sind voneinander stochastisch unabhängig.

Die Menge der betrachteten Farbkanäle werde im Weiteren mit K bezeichnet. Zum Beispiel ist $K = \{R, G, B\}$ für den RGB-Farbraum oder $K = \{Y, r, g\}$ für das Yrg-Farbmodell. Sei $x_{t,k}$ ein zum Zeitpunkt t in einem Kanal $k \in K$ auftretender Farbwert, sowie $\mu_{t,k}$ und $\sigma_{t,k}$ der zu diesem Kanal gehörende Mittelwert und die Standardabweichung der Normalverteilung des beobachteten Pixel an der Stelle (i, j). Dieser Pixel wird dann als zum Vordergrund gehörend klassifiziert, falls für alle betrachteten Kanäle $k \in K$ gilt:

$$|x_{t,k} - \mu_{t,k}| > 2.5\sigma_{t,k}$$

Abschließend ermöglicht der Einsatz eines schließenden morphologischen Filters auf den Vordergrundpixeln, Lücken zu schließen und so möglichst große zusammenhängende Vordergrundregionen zu erhalten.

Um auf zeitliche Veränderungen zu reagieren, werden die Mittelwerte und Standardabweichungen für alle betrachteten Farbkanäle wie folgt angepasst:

$$\mu_{t+1,k} = (1 - \alpha)\mu_{t,k} + \alpha(x_{t,k} - \mu_{t,k})$$

$$\sigma_{t+1,k} = (1 - \alpha)\sigma_{t,k} + \alpha(x_{t,k} - \mu_{t,k})^2$$

$\alpha \in [0, 1]$ bezeichnet dabei die Lernrate. Wie leicht ersichtlich ist, werden Vordergrundpixel umso schneller zum Hintergrund adaptiert je größer α ist, da das Gewicht der früher beobachteten Farbwerte dann umso kleiner wird. Die beiden Abbildungen 2.2 und 2.3 sollen beispielhaft das Ergebnis der Segmentierung illustrieren. Dabei sind Pixel, die als Vordergrundbereiche klassifiziert wurden, schwarz eingefärbt.

An den Beispielbildern wird auch der Vorteil der Verwendung des Yrg-Farbraums sichtbar, nämlich dass Schattenbereiche wie gewünscht wesentlich seltener zum Vordergrund gerechnet werden. Erkauft wird dieser Vorteil allerdings durch die Einschränkung, dass das Modell insgesamt unempfindlicher auf Farbänderungen wird und dadurch kleine Bewegungen von Sprechern nur sehr unzureichend erkannt werden können.

Abbildung 2.2: Segmentierungsergebnis des Frames 156, Bildfolge Seltzer, Kamera 1 im RGB-Farbraum

Abbildung 2.3: Segmentierungsergebnis des Frames 156, Bildfolge Seltzer, Kamera 1 im Yrg-Farbraum

2.3 Triangulation

Nachdem die Bilder der vier Kameras segmentiert wurden, hat man die Vordergrundbereiche aus den jeweiligen Perspektiven erhalten. Ziel ist es nun, daraus die möglichen Raumkoordinaten des Sprechers in 3D zu bestimmen, was durch die Kalibrierung der verwendeten Kameras ermöglicht wird.
Dabei wird ausgehend von den Sichtgeraden durch die optischen Zentren der vier Kameras auf die Silhouetten ein überbestimmtes lineares Gleichungssystem zum Schnitt der Geraden aufgestellt, das durch Minimierung des quadratischen Fehlers gelöst wird. Der interessierte Leser findet hierzu Details bei Focken [6].

Im Folgenden sollen jedoch zumindest die wichtigsten Schritte knapp erläutert werden. Zunächst wird für jeden Vordergrundbereich der Mittelpunkt an der höchsten Stelle, im Optimalfall die Position des Kopfes, berechnet, für den das Korrespondenzproblem anschließend gelöst werden muss. Darunter versteht man die richtige Zuordnung von zueinander gehörenden Vordergrundbereichen auf den verschiedenen Kamerabildern. Problematisch hierbei sind allerdings folgende Fälle:

- Die Silhouette des Sprechers zerfällt bei der Segmentierung in mehrere Teilsilhouetten.

- Zusammengehörende Vordergrundregionen können nicht in allen Perspektiven beobachtet werden.

- Durch Rauschen können zusätzliche nicht zum Sprecher gehörende Vordergrundbereiche entstehen.

Da a priori nicht entschieden werden kann, welche Zentroidteilmengen eine Korrespondenz formen, wird versucht für alle Teilmengen von Zentroiden auf verschiedenen Kamerabildern das entsprechende Gleichungssystem zu lösen. Anschließend werden Lösungen, deren quadratischer Fehler, das Residuum, einen Schwellwert übersteigt, verworfen. Die verbleibenden Lösungen werden im Folgenden als 3D-Hypothesen z_t bezeichnet.

Zudem kann der Lösungsraum durch die Forderung eingeschränkt werden, dass Lösungen mindestens aus $n \in \{2, 3, 4\}$ Kameraperspektiven Vordergrundbereiche zur Berechnung verwendet haben. Dies birgt den Vorteil, dass Bewegungen des Publikums, das am Rand des sichtbaren Bereichs sitzt, nicht weiterverwendet werden, da Zuschauer meist nur aus maximal zwei Perspektiven zu sehen sind. Der Sprecher hingegen kann im Normalfall aus allen vier Perspektiven beobachtet werden, so dass Lösungen, die mögliche 3D-Koordinaten des Vortragenden sind, nicht von dieser Einschränkung betroffen sind.

2.4 Filterung

Allgemein schätzen Filter den Zustand x_t eines Systems zum Zeitpunkt t auf der Grundlage einer Menge von Messungen $Z_t = \{z_0, \ldots, z_t\}$. Der Systemzustand $x_t = (x_s\ y_s\ z_s\ \dot{x}_s\ \dot{y}_s\ \dot{z}_s)^T$ ist in diesem Fall die Position des Sprechers in 3D-Koordinaten, sowie dessen Bewegungsgeschwindigkeit; die 3D-Hypothesen der Triangulation stellen die Messungen Z_t dar. Meist wurde in bisherigen Anwendungen, so auch bei Focken [6], ein Kalmanfilter verwendet, der allerdings den Nachteil hat, dass er alle Einflüsse nur als Normalverteilungen modelliert. Da dies insbesondere bei gestörten Eingabedaten nicht immer den realen Gegebenheiten entspricht, soll auch ein Partikelfilter nach Blake und Isard [10] zum Einsatz kommen.

2.4.1 Kalmanfilter

Im Wesentlichen verwenden Kalmanfilter zwei Modelle um den Systemzustand x_t über die Zeit zu propagieren, wobei alle auftretenden Einflussgrößen als normalverteilt angenommen werden.
Dieses sind zum einen das Systemmodell, das den vermuteten Aufenthaltsort x_{t+1} in Abhängigkeit von x_t modelliert, sowie das Beobachtungsmodell, das die Wahrscheinlichkeit von Messungen z_{t+1} in Abhängigkeit von x_{t+1} beschreibt. Details insbesondere zur mathematischen Herleitung der Propagierungsregeln können aus Russell [16] entnommen werden. Das verwendete Beobachtungsmodell lässt sich wie folgt formulieren:

$$x_{t+1} = A_t x_t + B_t u_t + w_t$$

wobei gelten soll:

$$x_t = \begin{pmatrix} x_s \\ y_s \\ z_s \\ \dot{x}_s \\ \dot{y}_s \\ \dot{z}_s \end{pmatrix}; \qquad A_t = A = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}; \qquad B_t = B = 0$$

w_t beschreibt das sogenannte Prozessrauschen, die Dynamik des Systems, im vorliegenden Fall die Beschleunigung des Sprechers durch eine mehrdimensionale Normalverteilung mit Mittelwert 0 und folgender Kovarianzmatrix:

$$Q_t = Q = \begin{pmatrix} q_{11}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & q_{22}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & q_{33}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & q_{44}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & q_{55}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & q_{66}^2 \end{pmatrix}$$

Das Beobachtungsmodell kann beschrieben werden durch:

$$z_t = Hx_t + v_t$$

Dabei hat H die folgende Form:

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

v_t, das als Maß für die Messgenauigkeit eingeführt wird, wird ebenfalls normalverteilt mit Mittelwert 0 und folgender Kovarianzmatrix beschrieben:

$$R_t = R = \begin{pmatrix} r_{11}^2 & 0 & 0 \\ 0 & r_{22}^2 & 0 \\ 0 & 0 & r_{33}^2 \end{pmatrix}$$

Da $z_{t+1} = \{z_{t+1}^1, \ldots, z_{t+1}^m\}$ meist eine Menge von m Messungen darstellt, muss entschieden werden, welche für das Beobachtungsmodell weiter verwendet werden soll. Dafür wird diejenige Messung z_{t+1}^i ausgewählt, die gemäß der euklidschen Norm den geringsten Abstand d zur letzten Schätzung hat.

$$d = argmin_{i=1\ldots m}(\left|\left|z_{t+1}^i - x_t\right|\right|_2)$$

Initialisiert wird der Statusvektor x_t anhand der Korrespondenz mit dem kleinsten Residuum der ersten Messung z_t und dem Geschwindigkeitsvektor 0. Für die Implementierung wurden im Übrigen die Methoden zur Kalmanfilterung aus der Open Source Computer Vision Library verwendet.

2.4.2 Partikelfilter

Bei Partikelfiltern wird die Aufenthaltswahrscheinlichkeit des zu verfolgenden Objekts nicht wie bei Kalmanfiltern durch eine geschlossene Gleichung mittels Normalverteilungen angegeben sondern durch eine endliche Menge diskreter Einzelhypothesen genähert. Insbesondere ermöglicht dies auch, statt einer einzelnen Person ohne wesentlich größeren Aufwand mehrere Personen zu verfolgen. Als weiterer Vorteil ergibt sich eine im Vergleich zu Kalmanfiltern wesentlich größere Robustheit gegenüber Störungen [10], die dadurch erreicht wird, dass mehrere mögliche Hypothesen gleichzeitig verfolgt werden.

Da sich diese Arbeit im Wesentlichen mit der Verwendung von Partikelfiltern für die Personenverfolgung in Seminarszenarien beschäftigt, sollen die verwendeten Methoden und Modelle in Kapitel 3 näher betrachtet werden.

3 Funktionsweise von Partikelfiltern

Partikelfilter begegnen dem Problem von Kalmanfiltern, nur unimodale Wahrscheinlichkeitshypothesen zu machen, mit der Idee des „factored sampling“, bei dem Wahrscheinlichkeitsdichten durch eine zufällig generierte Menge von „Partikeln“ repräsentiert werden. Diese Menge wird auf der Grundlage eines dynamischen Modells, das auch als Bewegungsmodell bezeichnet wird, propagiert und durch die Anwendung eines Beobachtungsmodells neu bewertet. Dadurch erhält man als Ergebnis einen Filter, der gegenüber Störungen wesentlich robuster ist. Im Wesentlichen stützt sich dieses Kapitel auf die Arbeit von Isard und Blake [10].

3.1 Notationen und Annahmen

Wie in Abschnitt 2.4 eingeführt, wird auch hier der Zustandsvektor mit $x_t = (x_s\ y_s\ z_s\ \dot{x}_s\ \dot{y}_s\ \dot{z}_s)^T$ bzeichnet. Die Historie der Messungen bis zum Zeitpunkt t soll $Z_t = \{z_0, \ldots, z_t\}$ genannt werden, wobei für jedes $i \in \{0, \ldots, t\}$ gelten soll:

$$z_i = \{z_t^1, \ldots, z_t^m\}$$

Die Vergangenheit des Zustandsvektors soll die Bezeichnung

$$X_t = \{x_0, \ldots, x_t\}$$

tragen.

Außerdem werden einige Annahmen bezüglich der Abhängigkeit von Messungen und Statusvektor gemacht werden, um die Modellierung zu vereinfachen. Dies ist zunächst die Forderung, dass die Bewegungsdynamik des Sprechers zum Zeitpunkt t nur vom Status x_{t-1} abhängt und die weitere Vergangenheit unberücksichtigt bleibt:

$$p(x_t|X_{t-1}) = p(x_t|x_{t-1})$$

Die Dynamik des Sprechers ist also bezüglich der Zeit eine Markovkette.

Für die Messungen wird angenommen, dass diese sowohl untereinander voneinander unabhängig sind, als auch im Hinblick auf den dynamischen Prozess:

$$p(Z_{t-1}, x_t|X_{t-1}) = p(x_t|X_{t-1}) \prod_{i=1}^{t-1} p(z_i|x_i)$$

$$p(Z_t|X_t) = \prod_{i=1}^{t} p(z_i|x_i)$$

3.2 Voranschreiten des Systemzustands

Wie bereits einleitend erwähnt, wird die Aufenthaltswahrscheinlichkeit $p(x_t|Z_t)$ des Sprechers durch eine Menge von diskreten Einzelhypothesen, den Partikeln approximiert, die über die Zeit propagiert werden. Die mathematische Gültigkeit dieses Vorgehens lässt sich unter Zuhilfenahme der Regel von Bayes nachvollziehen, nach der gilt:

$$p(x_t|Z_t) = k_t \; p(z_t|x_t)p(x_t|Z_{t-1})$$

Dabei wird $p(z_t|x_t)$ mit Hilfe des Beobachtungsmodells bestimmt. k_t ist ein von X_t unabhängiger Normierungsfaktor, der für die Maximierung von $p(x_t|z_t)$ unberücksichtigt und zu eins gesetzt werden kann. Die a-priori Wahrscheinlichkeit $p(x_t|Z_{t-1})$ lässt sich darüber hinaus in Abhängigkeit von der a-posteriori Wahrscheinlichkeitsdichte $p(x_{t-1}|Z_{t-1})$ ausdrücken durch:

$$p(x_t|Z_{t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|Z_{t-1})dx_{t-1}$$

Der Beweis hierzu findet sich im Anhang von [10]. Mit anderen Worten gesagt, lässt sich die Wahrscheinlichkeit des Systemzustands x_t in Abhängigkeit von den Messungen Z_{t-1} bis zum letzten Zeitpunkt $t-1$ dadurch gewinnen, dass man alle Hypothesen mit Statusvektor x_{t-1} mit den Auftrittswahrscheinlichkeiten $p(x_t|x_{t-1})$ des dynamischen Modells gewichtet. Eine weitere Gewichtung mit der Wahrscheinlichkeitsdichtefunktion $p(z_t|x_t)$ aus dem Beobachtungsmodell, die den Einfluss der Messungen z_t berücksichtigt, liefert dann die Schätzung des Statusvektors des Filters für das vorliegende System, also Geschwindigkeit und Aufenthaltsort des Sprechers. Abbildung 3.1, die aus dem Artikel von Isard und Blake [10] entnommen ist, soll diesen Sachverhalt nochmals bildlich darstellen.

3.3 Factored Sampling

Factored Sampling beschreibt nach Grenander et al. [8] die Möglichkeit die a-posteriori Wahrscheinlichkeitsdichte $p(x|z)$ durch eine Menge von „Samples“, die

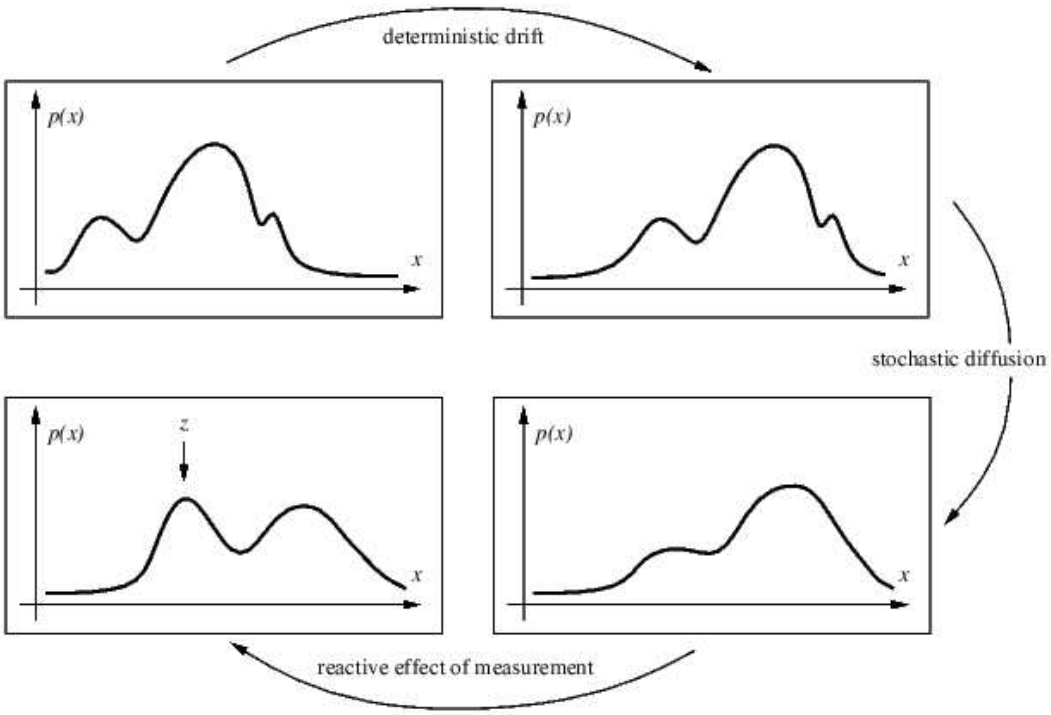

Abbildung 3.1: Voranschreiten des Systemzustands unter Berücksichtigung von Bewegungs- und Beobachtungsmodell für eindimensionalen Statusvektor x [10]

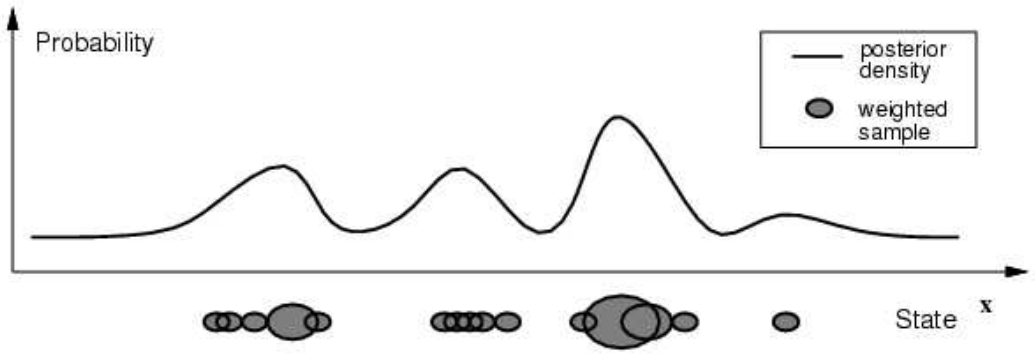

Abbildung 3.2: Näherung von $p(x|z)$ durch Samplemenge [10]

auch als „Partikel“ bezeichnet werden, zu nähern. Dabei wird zunächst die Samplemenge $s = \{s^{(1)}, \ldots, s^{(N)}\}$, eine Menge von Einzelhypothesen des Statusvektors, mit einer Verteilung auf Grundlage der a-priori Wahrscheinlichkeitsdichte $p(x)$ erzeugt.
Jedem Sample $s^{(i)}$ $(i \in \{1, \ldots, N\})$ wird zudem ein Gewicht

$$\pi_i = \frac{p(z|s^{(i)})}{\sum_{j=1}^{N} p(z|s^{(j)})}$$

zugeordnet. Durch Ziehen mit Zurücklegen entsprechend den Gewichten π_i als Wahrscheinlichkeiten erhält man die Samplemenge $\tilde{s}$, die $p(x|z)$ für hinreichend großes N nähert. Dies wird in Abbildung 3.2 veranschaulicht.

3.4 Der Condensation Algorithmus

Der Condensation Algorithmus [10] stellt eine iterative Form des factored sampling dar. Zu jedem Zeitpunkt t existiert eine mit $\pi_t = \{\pi_t^{(n)} | n = 1, \ldots, N\}$ gewichtete Samplemenge $s_t = \{s_t^{(n)} | n = 1, \ldots, N\}$, die die Wahrscheinlichkeitsdichte $p(x_t|Z_t)$ für den Aufenthaltsort des Sprechers widergibt. Zusätzlich wird für jeden Partikel i

$$c_t^{(i)} = \sum_{j=1}^{j=i} \pi_t^{(j)}$$

gespeichert, was es ermöglicht, den Resamplingschritt in einer Laufzeit von $O(n \log(n))$ statt in $O(n^2)$ durchzuführen.

Um die Schätzung für den Zeitpunkt $t + 1$ zu ermitteln wird wie folgt vorgegangen:

Erzeugung: Eine neue, noch ungewichtete Samplemenge s_{t+1} wird duch Ziehen mit Zurücklegen aus der alten Partikelmenge s_t erzeugt. Dabei wird eine auf $[0, 1]$ gleichverteilte Zufallszahl r bestimmt und durch binäre Suche dasjenige Partikel mit dem kleinsten j ausgewählt, für das gilt:

$$c_t^{(j)} \geq r$$

Bewegung: Die Partikel dieser Menge s_{t+1} werden dann, getrieben durch das Bewegungsmodell, deterministisch bewegt und anschließend mit Rauschen überlagert.

Messung: Die Gewichte $\pi_{t+1} = p(Z_{t+1}|x_{(t+1)} = s_{t+1}^{(n)})$ werden durch das Beobachtungsmodell bestimmt und abschließend normiert, so dass gilt:

$$\sum_{i=1}^{N} \pi_{t+1}^{(i)} = 1$$

Schätzung: Um schließlich den geschätzten Aufenthaltsort des Sprechers in 3D-Koordinaten zu bestimmen, wird der mit π_{t+1} gewichtete Mittelwert der Einzelhypothesen s_{t+1} gebildet:

$$\begin{pmatrix} x_{e,t+1} \\ y_{e,t+1} \\ z_{e,t+1} \end{pmatrix} = \sum_{i=1}^{N} \pi_{t+1}^{(i)} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} s_{t+1}^{(i)}$$

Die Initialisierung der Ortskoordinaten des Statusvektors geschieht im implementierten System entweder durch eine vorgegebene Normalverteilung der Form

$$p(x_0) = \frac{1}{\sqrt{2\pi\,|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

$$mit\ \mu = \begin{pmatrix} \mu_x \\ \mu_y \\ \mu_z \end{pmatrix}\ und\ \Sigma = \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix}$$

oder durch eine Mixtur aus Normalverteilungen des a-priori Aufenthaltsmodells, auf das im Abschnitt 3.6.2 noch näher eingegangen werden wird. Der initiale Geschwindigkeitsvektor wird in beiden Fällen zu 0 gesetzt.

3.5 Bewegungsmodell

Das Bewegungsmodell beschreibt die angenommene deterministische Bewegung des Sprechers als Wahrscheinlichkeitsdichte $p(x_t|x_{t-1})$. Im Trackingssystem dieser

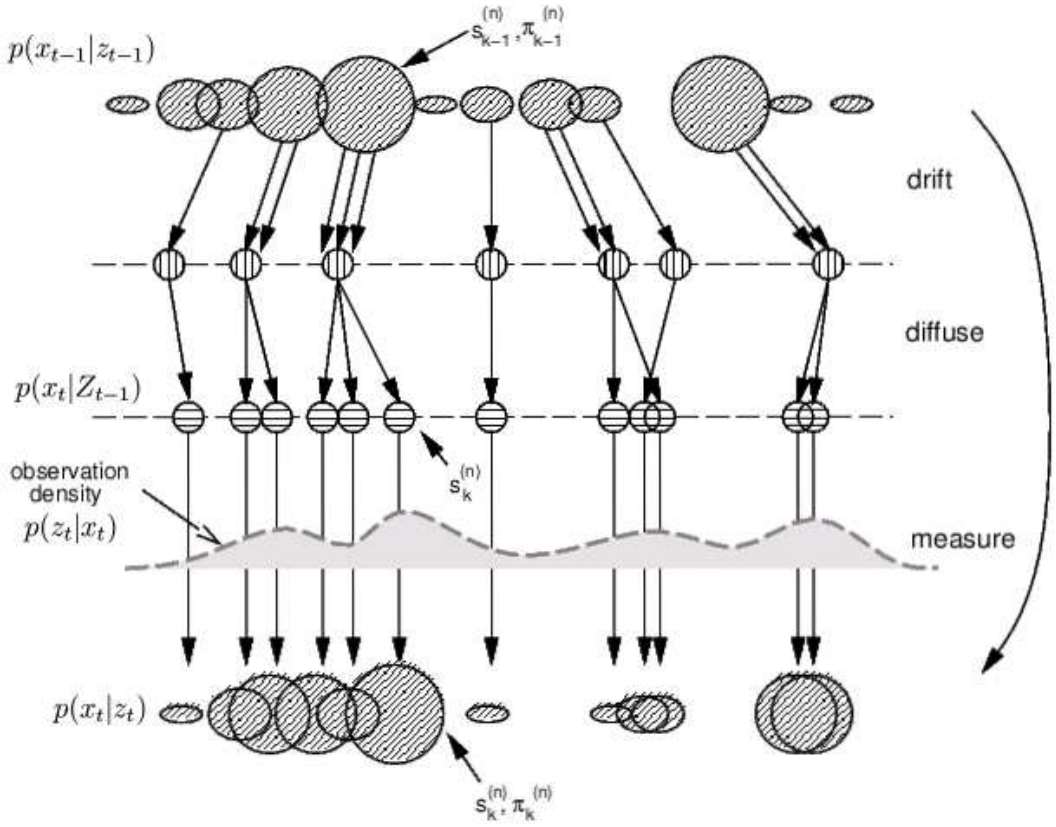

Abbildung 3.3: Ablauf einer Iteration des Condensation Algorithmus [10]

Arbeit wird ein simples lineares Bewegungsmodell verwendet, das heißt es wird angenommen, dass sich die Geschwindigkeit des Sprechers über die Zeit nicht ändert. Allerdings wird dazu eine normalverteilte Beschleunigung als Rauschen addiert, um auf Geschwindigkeitsänderungen zu reagieren. Dadurch werden Partikel mit der falschen Geschwindigkeit so lange propagiert, bis sie aufgrund schlechter Messwerte ein zu niedriges Gewicht $\pi_t^{(i)}$ erhalten und nicht mehr in die Partikelmenge s_{t+1} übernommen werden. Mathematisch lässt sich dies wie folgt ausdrücken:

$$s_{t+1}^{(i)} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} s_t^{(i)} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ a_x \\ a_y \\ a_z \end{pmatrix}$$

wobei a_d ($d \in \{x, y, z\}$) eine nach $\mathcal{N}(0, \sigma_d^2)$ verteilte Zufallszahl ist.

Problematisch für dieses Modell sind insbesondere Bewegungen des Sprechers, bei denen er abrupt beschleunigt oder seine Richtung wechselt. In diesem Fall wird die bisherige Bewegung nämlich solange fortgesetzt, bis die Partikel mit falschem Geschwindigkeitsvektor „absterben“ und Partikel in der näheren Umgebung von Messungen wieder stärkeren Einfluss auf die Schätzung haben.

Um unrealistische Bewegungen auszuschließen, wird außerdem ein Raummodell angewandt, das Partikel, die sich aus dem *Smartroom* entfernen wollen, mit Geschwindigkeit 0 an der entsprechenden Wand festhält.

3.6 Beobachtungsmodell

Das Beobachtungsmodell ist zuständig für die Neugewichtung der Samplemenge s_t unter Zuhilfenahme der Messdaten z_t. In dieser Arbeit soll zunächst ein Modell evaluiert werden, das Gewichte π_t einzig und allein auf der Grundlage des euklidischen Abstands im dreidimensionalen Raum berechnet. Im zweiten Abschnitt wird zusätzliches Wissen über die a-priori Aufenthaltswahrscheinlichkeit des Sprechers aus gelernten Daten mit eingebracht. Dadurch soll verhindert werden, dass Messungen, die aus der Bewegung von Zuhörern resultieren, fälschlicherweise für Messungen des Sprechers gehalten werden. Für beide Modelle gilt jedoch: Liegen keine Messungen vor oder wird allen Samples i das neue Gewicht $\pi_t^{(i)} = 0$ zugeordnet, dann sollen im nächsten Zeitschritt alle Samples mit gleicher Wahrscheinlichkeit zur neuen Samplemenge s_{t+1} beitragen:

$$\forall i : \pi_t^{(i)} = \frac{1}{N}$$

3.6.1 Ohne a-priori Aufenthaltswahrscheinlichkeiten

Falls z_t aus mehreren Messungen besteht, so muss zunächst entschieden werden, welche für die Neubewertung eines Partikels $\pi_t^{(i)}$ herangezogen wird. Am zweckmäßigsten erscheint es diejenige Messung für den Partikel $s_t^{(i)}$ aus $z_t = \{z_t^1, \ldots, z_t^M\}$ auszuwählen, deren euklidischer Abstand d_m minimal ist. Es wird also diejenige Messung m mit Abstand d_{min} gewählt, für die gilt:

$$d_{min,i} = argmin_{m=1,\ldots,M} \left(\left\| z_t^m - \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} s_t^{(i)} \right\|_2 \right)$$

Daraus errechnet sich das neu zugeordnete Gewicht $\pi_t^{(i)}$ auf Grundlage einer Normalverteilung $\mathcal{N}(0, \sigma_o^2)$:

$$\pi_t^{(i)} = p_{ohne\ a-priori\ WKD} = \frac{1}{\sqrt{2\pi}\sigma_o} e^{-\frac{1}{2}\frac{d_{min,i}^2}{\sigma_o^2}}$$

3.6.2 Mit a-priori Aufenthaltswahrscheinlichkeiten

Da oben genanntes Beobachtungsmodell sehr anfällig ist gegenüber Störungen, die aus Bewegungen des Publikums zu einem Zeitpunkt, an dem sich der Sprecher nicht bewegt, resultieren wird an dieser Stelle ein weiteres Beobachtungsmodell eingeführt. Dieses basiert darauf, dass eine Mischung von k Normalverteilungen für die erwarteten Positionen des Sprechers auf anderen Bildfolgen gelernt wurde. Als Lernalgorithmus für die Parameter wird der Expectation Maximization Algorithmus verwendet, der durch die vorherige Anwendung des k-means Clusteralgorithmus initialisiert wird. Man erhält also

$$p_o(x) = \sum_{i=1}^{k} \delta_i * w_i * \frac{1}{\sqrt{2\pi\,|\Sigma_i|}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}$$

$$\text{mit } x = \begin{pmatrix} x_x \\ x_y \\ x_z \end{pmatrix};\ \mu_i = \begin{pmatrix} \mu_{x,i} \\ \mu_{y,i} \\ \mu_{z,i} \end{pmatrix} \text{ und } \Sigma_i = \begin{pmatrix} \sigma_{x,i}^2 & 0 & 0 \\ 0 & \sigma_{y,i}^2 & 0 \\ 0 & 0 & \sigma_{z,i}^2 \end{pmatrix}$$

$$\delta_i = \begin{cases} 1 & \text{wenn } \forall r : |x_r - \mu_{r,i}| \leq 3 * \sigma_{r,i} \text{ wobei } r \in \{x, y, z\} \\ 0 & \text{sonst} \end{cases}$$

wobei w_i das Gewicht der jeweiligen Normalverteilung bezeichnet. Für die Neugewichtung der Samples soll dann gelten:

$$\pi_t^{(i)} = p_{ohne\ a-priori\ WKD} * p_o \left(\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} s_t^{(i)} \right)$$

4 Experimentelle Ergebnisse

Im folgenden Kapitel soll nun das implementierte Personenverfolgungssystem experimentell evaluiert werden. Dazu wurde in den verwendeten Seminardaten die Nase-/Mundregion in mehreren Perspektiven manuell markiert und die Position des Sprechers durch das oben beschriebene Triangulationsverfahren bestimmt. Die dadurch erhaltenen 3D-Koordinaten $(x_{Truth,t}\ y_{Truth,t}\ z_{Truth,t})^T$ wurden ebenso wie die Schätzungen des Trackingsystems $(x_{e,t}\ y_{e,t}\ z_{e,t})^T$ auf den Boden projiziert. Abschließend wurde der euklidische Abstand zwischen der Projektion des wirklichen Aufenthaltsortes und der geschätzten Position berechnet. Dieser wird im Weiteren als Fehler e_t bezeichnet und soll als Grundlage der Bewertung dienen.

$$P_{Truth,t} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_{Truth,t} \\ y_{Truth,t} \\ z_{Truth,t} \end{pmatrix}$$

$$P_{e,t} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_{e,t} \\ y_{e,t} \\ z_{e,t} \end{pmatrix}$$

$$e_t = ||P_{Truth,t} - P_{e,t}||_2$$

Zur Bewertung sollen zwei Videofolgen herangezogen werden. Die Videobildfolge „STZ“ mit einer Länge von 9000 Frames unterscheidet sich dabei von der Bildfolge „MYR“, die 10000 Einzelbilder umfasst, durch einen veränderten Farbabgleich sowie dadurch, dass wesentlich mehr Zuschauer auf den Kamerabildern zu sehen sind, da diese näher am Sprecher sitzen. Dadurch werden wesentlich mehr Bewegungen im Zuschauerraum registriert, wohingegen in der ersten Bildfolge die Zuschauer nur am äußersten Rand sichtbar sind. In den Abbildungen 4.1 und 4.2 ist jeweils ein Frame jeder Bildfolge beispielhaft dargestellt.

4.1 Farbraumvergleich RGB - Yrg

Zunächst soll ein Vergleich der Trackingergebnisse durchgeführt werden, bei dem der Farbraum von RGB zu Yrg variiert wird. Außerdem wird das einfache Beobachtungsmodell ohne gelernte a-priori Aufenthaltswahrscheinlichkeiten aus Abschnitt

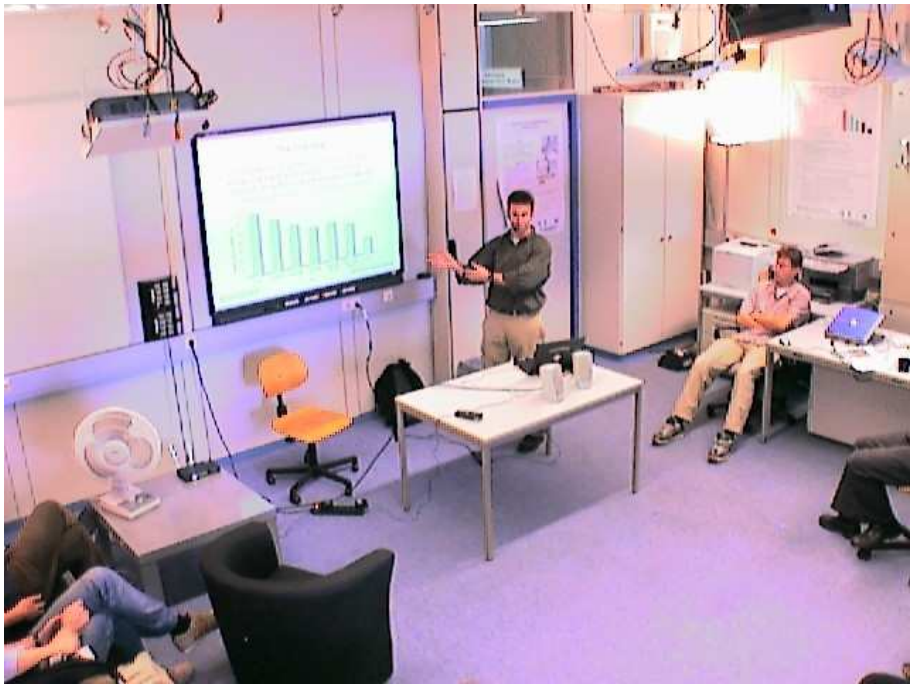

Abbildung 4.1: Frame 9975 der Sequenz „STZ“ aus der Sicht von Kamera 1

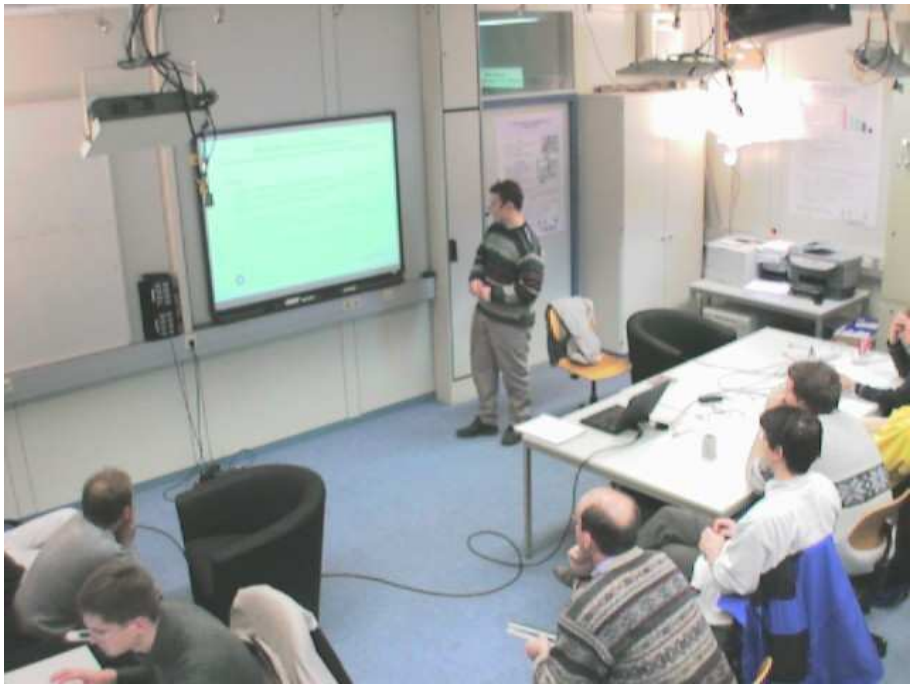

Abbildung 4.2: Frame 8345 der Sequenz „MYR“ aus der Sicht von Kamera 1

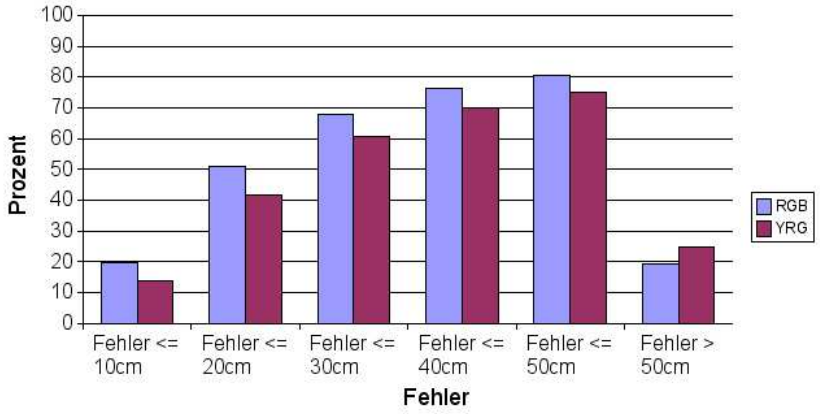

Abbildung 4.3: Trackingergebnisse auf der Sequenz „STZ“ in den Farbräumen RGB und Yrg

3.6.1 verwendet und es wird der Zentroid, der Mittelpunkt des Sprechers, verfolgt. Eine Übersicht über alle verwendeten Parameter findet sich für den interessierten Leser im Anhang A. Für die Folge „STZ“ ergibt sich dabei für den RGB Farbraum folgendes Ergebnis:

Tabelle 4.1: Fehler auf der Sequenz „STZ“ für den RGB-Farbraum

Frames mit $e_t \leq 10cm$	19,84%
Frames mit $e_t \leq 20cm$	50,94%
Frames mit $e_t \leq 30cm$	67,73%
Frames mit $e_t \leq 40cm$	76,19%
Frames mit $e_t \leq 50cm$	80,62%
Frames mit $e_t > 50cm$	19,38%
Durchschnittlicher Fehler $\bar{e_t}$	47,36cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	18,43cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	167,74cm

Die Berechnung mit Hilfe der Yrg-Kanäle liefert ein etwas schlechteres Ergebnis:

Tabelle 4.2: Fehler auf der Sequenz „STZ“ für den Yrg-Farbraum

Frames mit $e_t \leq 10cm$	13,91%
Frames mit $e_t \leq 20cm$	41,67%
Frames mit $e_t \leq 30cm$	60,82%
Frames mit $e_t \leq 40cm$	69,96%
Frames mit $e_t \leq 50cm$	75,02%
Frames mit $e_t > 50cm$	24,98%
Durchschnittlicher Fehler $\bar{e_t}$	51,33cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	20,17cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	144,92cm

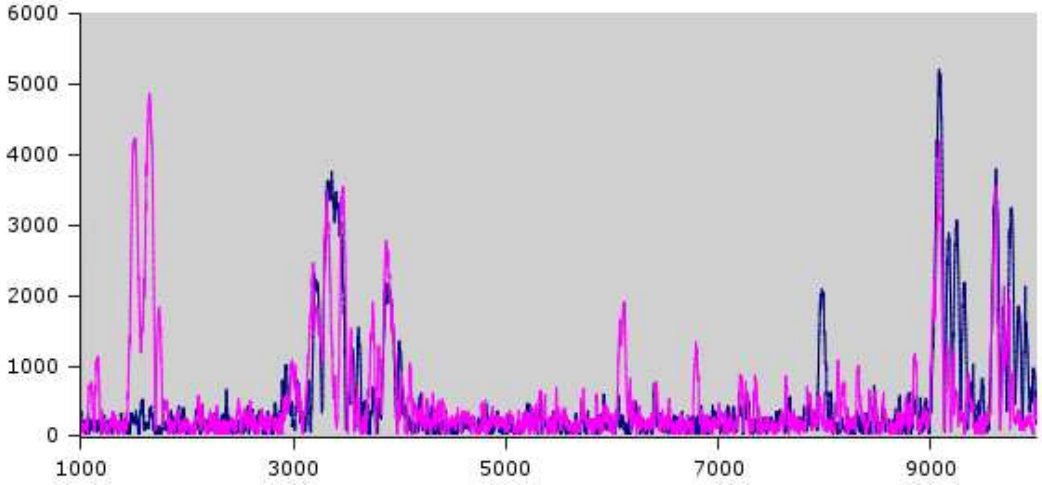

Abbildung 4.4: Fehler auf der Sequenz „STZ“ in mm aufgetragen über der der Framenummer; in rot Yrg, in blau RGB

Betrachtet man Abbildung 4.4 genauer, sind insbesondere die Fehler zwischen Frame 3000 und 4000 sowie zwischen 9000 und 10000 auffällig. Dabei treten zwei Problemfälle auf. Dies ist im ersten Fall eine durch das Bild laufende Person wie auf Abbildung 4.5 zu sehen ist. Im zweiten Fall wird die Verfolgung dadurch verfälscht, dass der Sprecher längere Zeit still steht und dadurch zum Hintergrund adaptiert wird, aber gleichzeitig Bewegung im Publikum stattfindet; in diesem Fall durch Signalisierung eines Handzeichens (Abbildung 4.6). Dabei kennzeichnen die roten Punkte mit blauen Umrandungen die segmentierten Bereiche mit Mittelpunkt, gelbe Punkte die 3D-Hypothesen, die ins Bild zurückprojiziert wurden und der grüne Punkt die Schätzung des Aufenthaltsorts des Partikelfilters.

Die erhöhte Fehlerrate zwischen Frame 1000 und 2000 bei der Segmentierung mit Hilfe des Yrg-Modells lässt sich durch die verminderte Empfindlichkeit des Modells gegenüber kleinen Bewegungen erklären.

Bei der zweiten Bildfolge fällt der Einfluss der Zuschauer umso mehr ins Gewicht, da diese hier näher am Sprecher sitzen und entsprechend mehr Bewegung zu sehen ist. Folglich ist zu erwarten, dass der Schätzungsfehler größer wird, was auch an dessen Größe und Häufigkeit in Abbildung 4.8 ersichtlich ist. Allerdings zeigt sich auch hier, dass das RGB-Modell dem Yrg-Modell überlegen ist. Das Yrg kann auf dieser Bildfolge nur in ganz wenigen Frames überhaupt Ergebnisse liefern, was vermutlich mit dem unterschiedlichen Farbabgleich der Kameras zusammen hängt. Da dies ein Tracking unmöglich macht, werden an dieser Stelle nur die Ergebnisse für das RGB-Modell genannt, auch wenn für dieses auf Grund des hohen Störeinflusses und mangelnder Bewegung des Sprechers die Spur häufig verloren geht:

Abbildung 4.5: Frame 3268 der Sequenz „STZ“: Zweite Person läuft durchs Sichtfeld aller Kameras

Abbildung 4.6: Frame 9049 der Sequenz „STZ“: Still stehender Sprecher bei Bewegung im Publikum

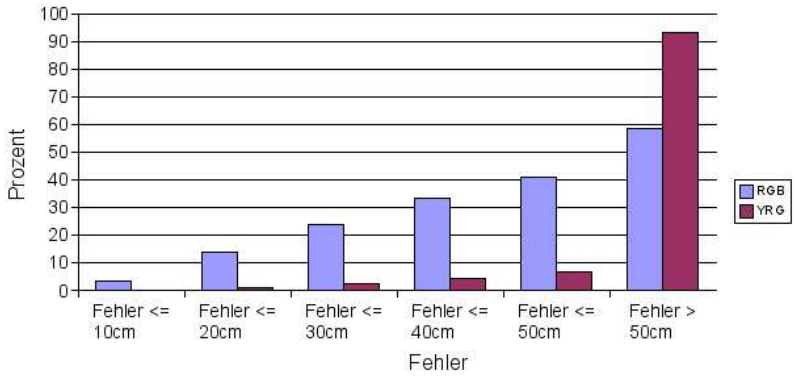

Abbildung 4.7: Trackingergebnisse auf der Sequenz „MYR“ in den Farbräumen RGB und Yrg

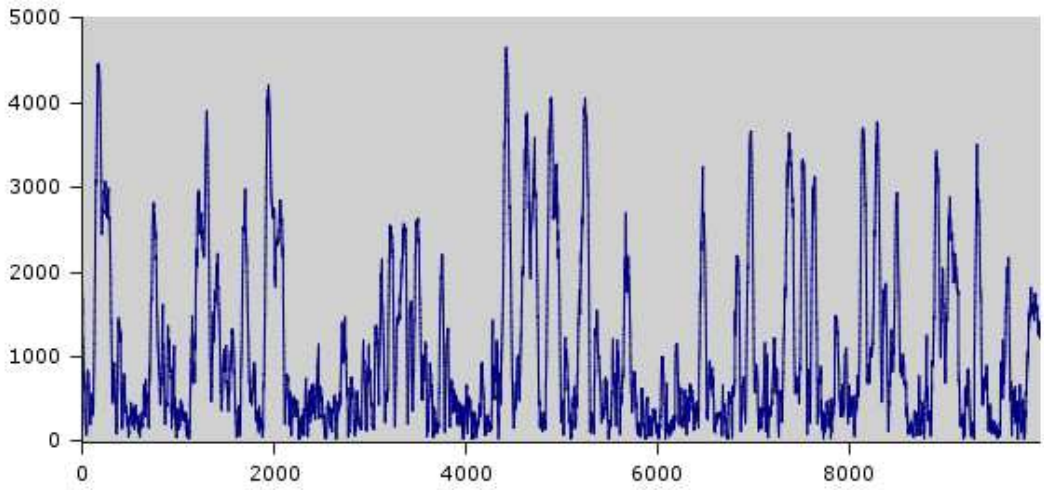

Abbildung 4.8: Fehler auf der Sequenz „MYR“ in mm aufgetragen über der der Framenummer

Tabelle 4.3: Fehler auf der Sequenz „MYR“ für den RGB-Farbraum

Frames mit $e_t \leq 10cm$	3,63%
Frames mit $e_t \leq 20cm$	13,87%
Frames mit $e_t \leq 30cm$	24,02%
Frames mit $e_t \leq 40cm$	33,54%
Frames mit $e_t \leq 50cm$	41,25%
Frames mit $e_t > 50cm$	58,75%
Durchschnittlicher Fehler $\bar{e_t}$	103,88cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	27,00cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	157,86cm

Auf Grund dieser Ergebnisse wird in den folgenden Experimenten das RGB-Modell zum Einsatz kommen.

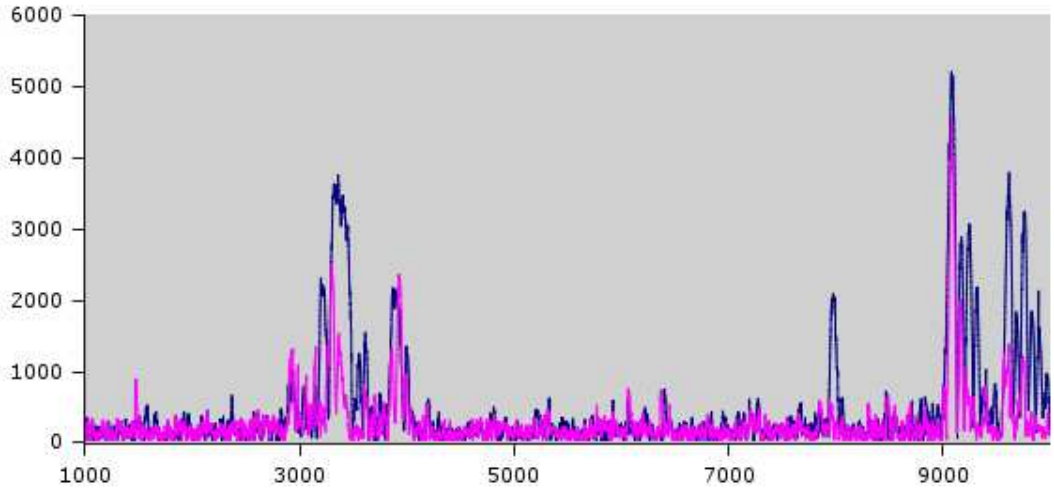

Abbildung 4.9: Fehler auf der Sequenz „STZ“ in mm aufgetragen über der der Framenummer; in rot mit Einschränkung der Messung, in blau ohne Einschränkung

4.2 Mindestzahl beobachtender Kameras

In diesem Abschnitt soll versucht werden den Störeinfluss der Zuschauer dadurch zu reduzieren, dass es nur noch zu gültigen 3D-Hypothesen kommt für den Fall, dass diese aus mindestens drei Perspektiven bei der Segmentierung zu beobachten sind. Allerdings sind weiterhin Störungen zu erwarten, da manche Zuschauer aus allen drei Perspektiven zu sehen sind. In der graphischen Veranschaulichung (Abbildung 4.9) kann man jedoch erkennen, dass sich das Problem insbesondere an den im vorherigen Abschnitt behandelten Stellen sichtbar verbessert.

Tabelle 4.4: Fehler auf der Sequenz „STZ“ für Mindestzahl beobachtender Kameras

Frames mit $e_t \leq 10cm$	24,10%
Frames mit $e_t \leq 20cm$	59,61%
Frames mit $e_t \leq 30cm$	78,26%
Frames mit $e_t \leq 40cm$	85,07%
Frames mit $e_t \leq 50cm$	88,58%
Frames mit $e_t > 50cm$	11,42%
Durchschnittlicher Fehler $\bar{e_t}$	27,95cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	17,22cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	111,18cm

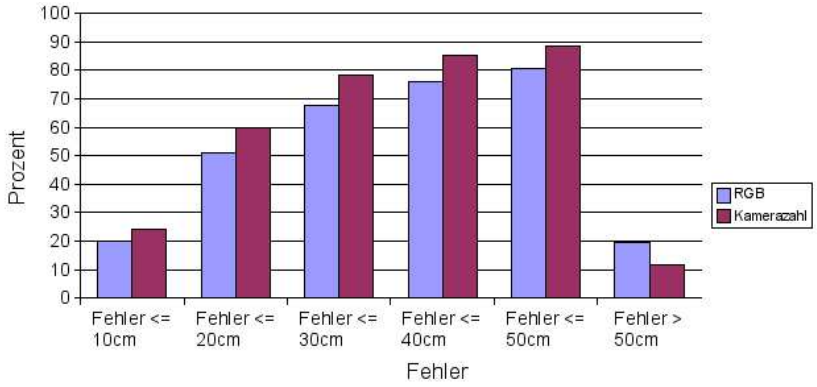

Abbildung 4.10: Trackingergebnisse auf der Sequenz „STZ“ mit Einschränkung durch Mindestzahl beobachtender Kameras

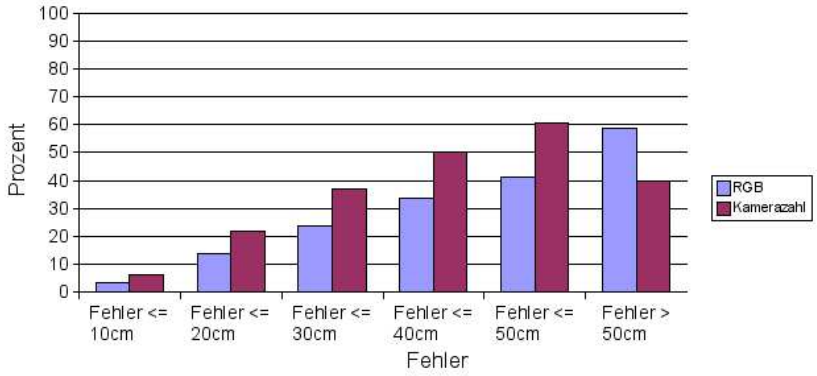

Abbildung 4.11: Trackingergebnisse auf der Sequenz „MYR“ mit Einschränkung durch Mindestzahl beobachtender Kameras

Auch für die zweite Bildfolge „MYR“ ergibt sich eine deutliche Verbesserung:

Tabelle 4.5: Fehler auf der Sequenz „MYR“ für Mindestzahl beobachtender Kameras

Frames mit $e_t \leq 10cm$	6,37%
Frames mit $e_t \leq 20cm$	21,83%
Frames mit $e_t \leq 30cm$	36,85%
Frames mit $e_t \leq 40cm$	50,03%
Frames mit $e_t \leq 50cm$	60,38%
Frames mit $e_t > 50cm$	39,62%
Durchschnittlicher Fehler $\bar{e_t}$	69,61cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	26,00cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	136,07cm

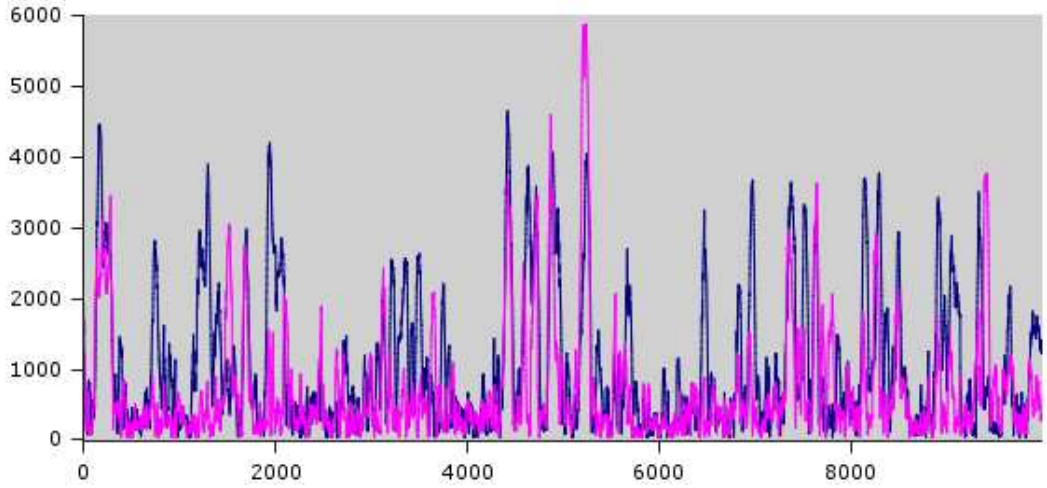

Abbildung 4.12: Fehler auf der Sequenz „MYR“ in mm aufgetragen über der Framenummer; in rot mit Einschränkung der Messung, in blau ohne Einschränkung

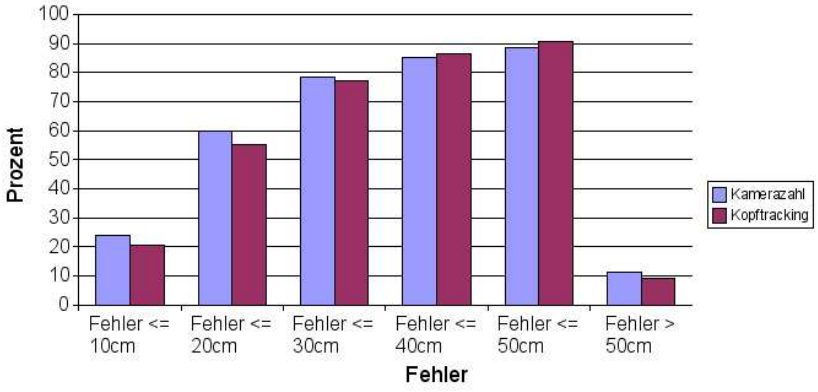

Abbildung 4.13: Trackingergebnisse auf der Sequenz „STZ“ bei Verfolgung des Kopfes

4.3 Verfolgung des Kopfes

In der Diplomarbeit von Focken [6] wird zudem vorgeschlagen, den Kopf statt des Zentroids zu verwenden, um die 3D-Hypothesen zu bestimmen. Dies soll an dieser Stelle ebenfalls evaluiert werden. Zu diesem Zweck wird nun also der Mittelpunkt der höchsten Position in segmentierten Vordergrundbereichen als Kopf angenommen. Diese Annahme ist natürlich nur für den Fall richtig, dass das Abbild des Sprechers nicht teilweise zum Hintergrund adaptiert wurde und somit eventuell in mehrere Teilbereiche zerfallen ist. Für die aufgezeichneten Daten der Bildfolge „STZ“ ergeben sich folgende Werte:

Tabelle 4.6: Fehler auf der Sequenz „STZ“ bei der Verfolgung des Kopfes

Frames mit $e_t \leq 10cm$	20,56%
Frames mit $e_t \leq 20cm$	54,96%
Frames mit $e_t \leq 30cm$	77,12%
Frames mit $e_t \leq 40cm$	86,38%
Frames mit $e_t \leq 50cm$	90,72%
Frames mit $e_t > 50cm$	9,28%
Durchschnittlicher Fehler $\bar{e_t}$	26,12cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	18,66cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	99,05cm

Offensichtlich hat sich der Anteil der Frames mit niedrigem Fehler minimal verringert, wohingegen allerdings der durchschnittliche Fehler pro Frame geringer wurde, weshalb in den weiteren Berechnungen weiter der höchste Punkt zur Berechnung der 3D-Hypothesen verwendet werden soll.

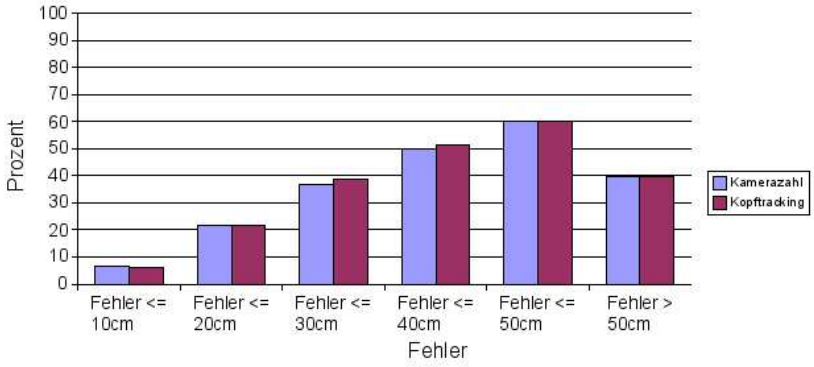

Abbildung 4.14: Trackingergebnisse auf der Sequenz „MYR“ bei Verfolgung des Kopfes

Das Ergebnis für den zweiten Seminarvortrag „MYR“:

Tabelle 4.7: Fehler auf der Sequenz „MYR“ bei der Verfolgung des Kopfes

Frames mit $e_t \leq 10cm$	6,13%
Frames mit $e_t \leq 20cm$	21,49%
Frames mit $e_t \leq 30cm$	38,60%
Frames mit $e_t \leq 40cm$	51,27%
Frames mit $e_t \leq 50cm$	60,29%
Frames mit $e_t > 50cm$	39,71%
Durchschnittlicher Fehler $\bar{e_t}$	70,22cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	25,67cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	137,83cm

Es zeigt sich also hier genauso wie in der ersten Bildreihe kaum ein Unterschied in der Verteilung der Fehlerwerte, jedoch eine Verbesserung des durchschnittlichen Fehlers. Erklärt werden kann dies wohl dadurch, dass die segmentierten Bereiche einer Person zu oft in mehrere kleine zerfallen.

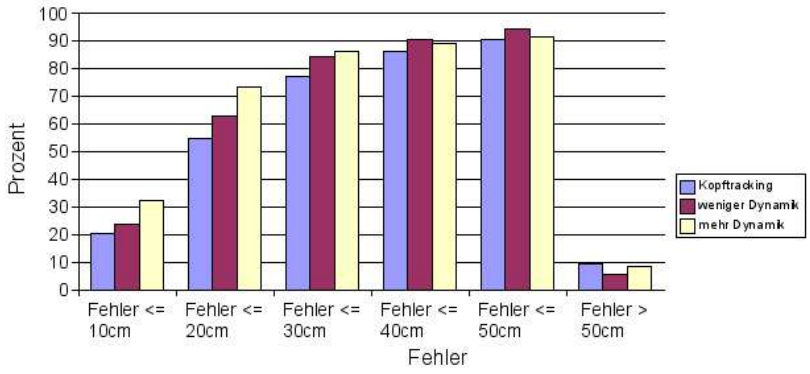

Abbildung 4.15: Trackingergebnisse auf der Sequenz „STZ“ mit unterschiedlicher Parametrisierung des dynamischen Modells

4.4 Veränderung der Objektdynamik

Als nächstes soll überprüft werden, inwiefern die Parameter des dynamischen Modells gut gewählt sind. Dazu wird die zufällige Beschleunigung des Modells aus Abschnitt 3.5 jeweils von $(0\ 0\ 0\ \mathcal{N}(0,(4\frac{m}{sf})^2)\ \mathcal{N}(0,(4\frac{m}{sf})^2)\ \mathcal{N}(0,(2\frac{m}{sf})^2))^T$ auf $(0\ 0\ 0\ \mathcal{N}(0,(8\frac{m}{sf})^2)\ \mathcal{N}(0,(8\frac{m}{sf})^2)\ \mathcal{N}(0,(4\frac{m}{sf})^2))^T$ erhöht oder auf $(0\ 0\ 0\ \mathcal{N}(0,(2\frac{m}{sf})^2)\ \mathcal{N}(0,(2\frac{m}{sf})^2)\ \mathcal{N}(0,(1\frac{m}{sf})^2))^T$ vermindert. Daraus ergeben sich für die verwendeten Bildfolgen folgende Ergebnisse:

Tabelle 4.8: Fehler auf der Sequenz „STZ“ bei Variation der dynamischen Parameter

Bildfolge „STZ“		
Fehler	**verminderte Beschleunigung**	**erhöhte Beschleunigung**
Frames mit $e_t \leq 10cm$	23,91%	32,34%
Frames mit $e_t \leq 20cm$	62,89%	73,31%
Frames mit $e_t \leq 30cm$	84,60%	86,24%
Frames mit $e_t \leq 40cm$	90,77%	89,44%
Frames mit $e_t \leq 50cm$	94,46%	91,37%
Frames mit $e_t > 50cm$	5,54%	8,63%
Durchschnittlicher Fehler $\bar{e_t}$	20,99cm	23,41cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	17,24cm	142,63cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	84,66cm	120,23cm

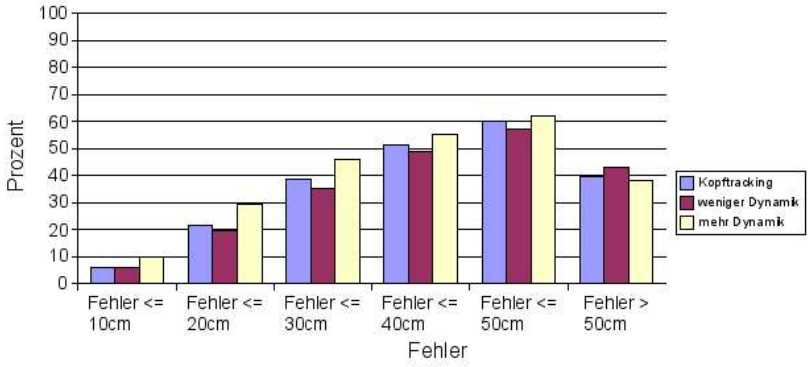

Abbildung 4.16: Trackingergebnisse auf der Sequenz „MYR“ mit unterschiedlicher Parametrisierung des dynamischen Modells

Tabelle 4.9: Fehler auf der Sequenz „MYR“ bei Variation der dynamischen Parameter

Bildfolge „MYR“		
Fehler	**verminderte Beschleunigung**	**erhöhte Beschleunigung**
Frames mit $e_t \leq 10cm$	5,93%	9,94%
Frames mit $e_t \leq 20cm$	19,51%	29,24%
Frames mit $e_t \leq 30cm$	35,49%	45,84%
Frames mit $e_t \leq 40cm$	49,04%	55,15%
Frames mit $e_t \leq 50cm$	57,05%	61,88%
Frames mit $e_t > 50cm$	42,95%	38,12%
Durchschnittlicher Fehler $\bar{e_t}$	65,70cm	80,68cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	25,76cm	22,40cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	118,76cm	175,28cm

Für den Seminarvortrag „STZ“ ergibt sich für die verminderte Beschleunigung ein etwas niedrigerer Durchschnittsfehler bei ebenfalls besserer Verteilung der Fehler, beim zweiten Seminarvortrag ergibt sich hingegen wenig Veränderung für niedrigere Dynamik. Für höhere Dynamik verbessert sich insbesondere die Genauigkeit, was daran liegt, dass der Bewegungsanteil in dieser Bildfolge wesentlich höher ist als im Seminarvortrag „STZ“.

4.5 Verwendung eines a-priori Aufenthaltsmodells

Da allerdings weiterhin die aus Bewegungen im Zuschauerraum resultierenden Probleme bestehen, soll deren Einfluss nun durch den Einsatz des in Abschitt 3.6.2 beschriebenen a-priori Aufenthaltsmodells reduziert werden. Dabei müssen allerdings auch einige sukzessive Abwandlungen vorgenommen werden, um den gewünschten Erfolg zu erzielen:

- Beschränkung des Aufenthaltsmodells auf x- und y-Koordinaten, da die Vordergrundbereiche des Sprechers zu oft in mehrere kleine zerfallen, wodurch auch die Hypothesen nicht mehr auf Kopfhöhe sind, auf der sich aber die gelernten Wahrscheinlichkeiten durch Markieren der Nase-/Mundregion befinden

- Verringerung des für 3D-Hypothesen erlaubten Bereichs auf Punkte innerhalb der 2σ-Umgebungen (vorher 3σ-Umgebungen) der Normalverteilungen

Tabelle 4.10: Fehler auf der Sequenz „STZ“ bei Einsatz eines a-priori Aufenthaltsmodells

Bildfolge „STZ“			
Fehler	3σ **im 3D**	3σ **im 2D**	2σ **im 2D**
Frames mit $e_t \leq 10cm$	7,32%	28,88%	29,59%
Frames mit $e_t \leq 20cm$	23,18%	73,72%	74,48%
Frames mit $e_t \leq 30cm$	37,74%	92,99%	93,00%
Frames mit $e_t \leq 40cm$	50,60%	97,70%	97,66%
Frames mit $e_t \leq 50cm$	61,27%	98,84%	98,62%
Frames mit $e_t > 50cm$	38,73%	1,16%	1,38%
Durchschnittlicher Fehler $\bar{e_t}$	55,45cm	15,96cm	15,87cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	25,84cm	15,38cm	15,13cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	102,28cm	65,85cm	69,22cm

Betrachtet man das Diagramm in Abbildung 4.17 so stellt man leicht fest, dass der gewünschte Effekt erreicht werden konnte. Die Zuschauerbewegungen konnten in beiden Problemfällen sehr effektiv unterdrückt werden und das Verfolgungsergebnis dadurch deutlich gesteigert werden.

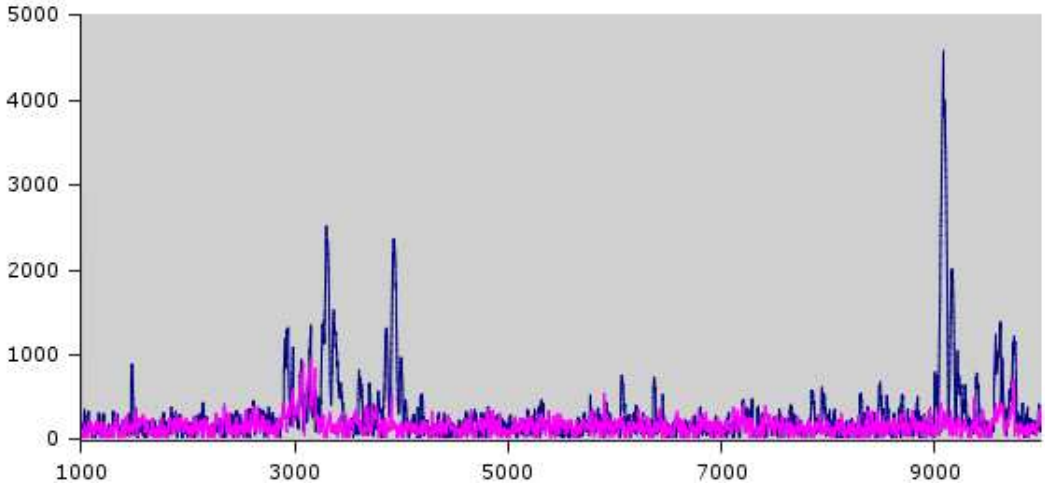

Abbildung 4.17: Fehler auf der Sequenz „STZ“ in mm aufgetragen über der der Framenummer; in rot mit a-priori Aufenthaltswahrscheinlichkeiten, in blau ohne a-priori Aufenthaltsmodell mit Einschränkung der Kamerazahl

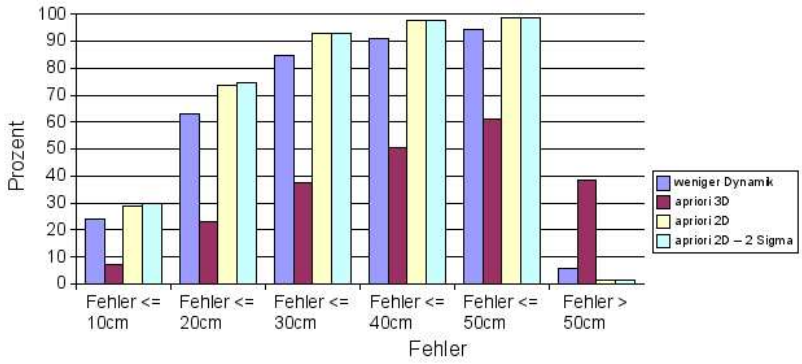

Abbildung 4.18: Trackingergebnisse auf der Sequenz „STZ“ bei Verwendung eines a-priori Aufenthaltsmodells

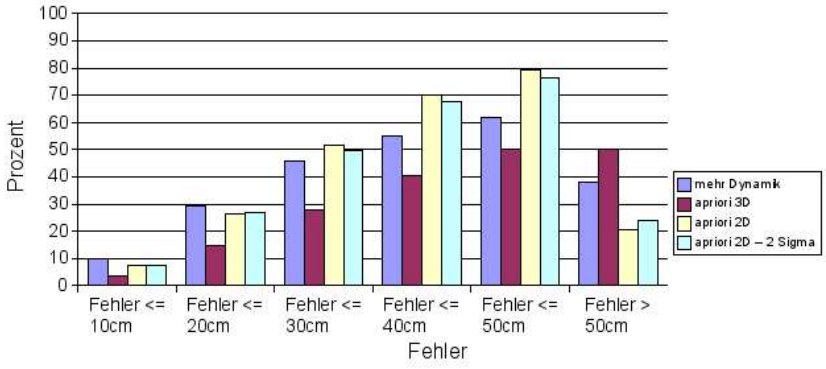

Abbildung 4.19: Trackingergebnisse auf der Sequenz „MYR“ bei Verwendung eines a-priori Aufenthaltsmodells

Tabelle 4.11: Fehler auf der Sequenz „MYR“ bei Einsatz eines a-priori Aufenthaltsmodells

Bildfolge „MYR“			
Fehler	3σ **im 3D**	3σ **im 2D**	2σ **im 2D**
Frames mit $e_t \leq 10cm$	3,65%	7,38%	7,63%
Frames mit $e_t \leq 20cm$	14,66%	26,52%	26,89%
Frames mit $e_t \leq 30cm$	28,10%	51,86%	49,71%
Frames mit $e_t \leq 40cm$	40,30%	69,95%	67,51%
Frames mit $e_t \leq 50cm$	50,00%	79,25%	76,16%
Frames mit $e_t > 50cm$	50,00%	20,75%	23,84%
Durchschnittlicher Fehler $\bar{e_t}$	90,50cm	49,77cm	58,36cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	27,87cm	25,45cm	25,21cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	153,12cm	142,66cm	164,26cm

Auch hier zeigt sich, wie aus Abbildung 4.20, dass der Einfluss der Zuschauerbewegungen deutlich reduziert werden kann. Auffällig bleibt allerdings der große Schätzfehler für die Frames 4000 bis 6000. Bei genauerem Betrachten der Bilddaten stellt man fest, dass sich der Sprecher hier vor der Tafel links der Projektionsfläche befindet. Problematisch ist hierbei, dass er dabei maximal von zwei Kameras beobachtet werden kann, was dazu führt, dass für den Zeitraum von etwa 1500 Frames kaum 3D-Hypothesen vorliegen und die Spur auf Grund des linearen Bewegungsmodells verloren geht beziehungsweise durch die stochastische Diffusion in eine beliebige Richtung abgelenkt werden kann.

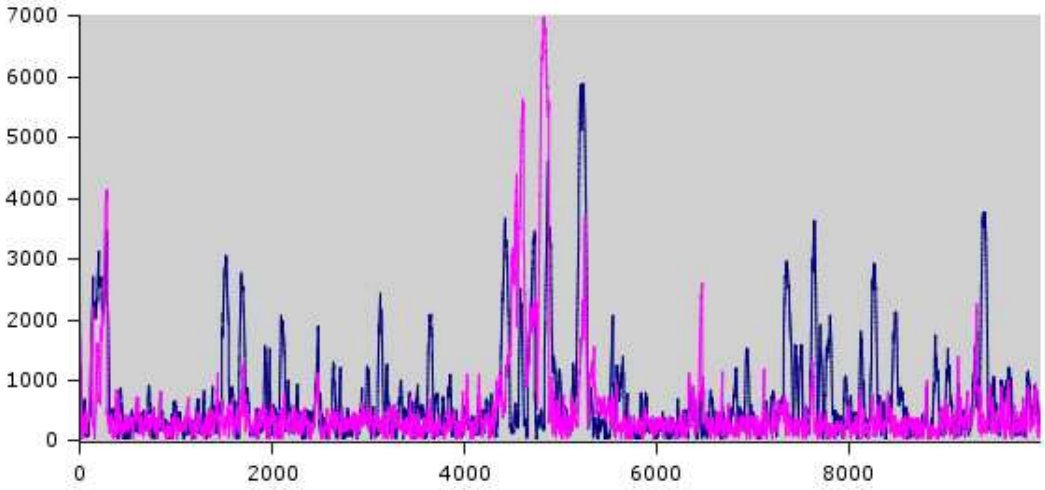

Abbildung 4.20: Fehler auf der Sequenz „MYR“ in mm aufgetragen über der der Framenummer; in rot mit a-priori Aufenthaltswahrscheinlichkeiten, in blau ohne a-priori Aufenthaltsmodell mit Einschränkung der Kamerazahl

Abbildung 4.21: Trackingergebnisse auf der Sequenz „STZ“ bei Erhöhung der Partikelzahl

4.6 Erhöhung der Partikelzahl

Als letzter Einflussfaktor auf das Trackingergebnis soll nun noch die Zahl der Partikel überprüft werden. Dazu wird die Zahl von bisher 500 Samples auf 2500 erhöht. Legt man den Vergleich von Zotkin [21] zu Grunde, dann sollte auch hieraus eine nochmalige Verbesserung des Ergebnisses resultieren.

Tabelle 4.12: Fehler auf der Sequenz „STZ“ bei Erhöhung der Partikelzahl

Frames mit $e_t \leq 10cm$	34,24%
Frames mit $e_t \leq 20cm$	82,98%
Frames mit $e_t \leq 30cm$	96,72%
Frames mit $e_t \leq 40cm$	98,74%
Frames mit $e_t \leq 50cm$	99,79%
Frames mit $e_t > 50cm$	0,21%
Durchschnittlicher Fehler $\bar{e_t}$	13,78cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	13,69cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	56,14cm

Bei Betrachtung der Tabelle fällt auf, dass vor allem die Genauigkeit durch die Erhöhung der Partikelzahl verbessert werden konnte, wie es bereits erwartet wurde.

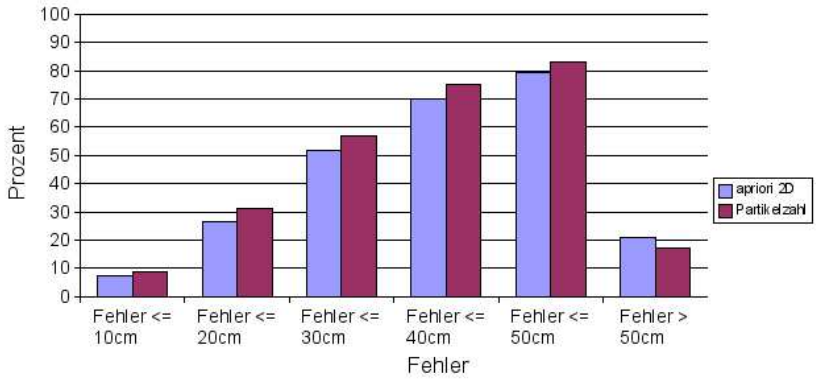

Abbildung 4.22: Trackingergebnisse auf der Sequenz „MYR“ bei Erhöhung der Partikelzahl

Ein ähnliches Bild zeigt sich auch für die zweite evaluierte Bildfolge „MYR“:

Tabelle 4.13: Fehler auf der Sequenz „MYR“ bei Erhöhung der Partikelzahl

Frames mit $e_t \leq 10cm$	8,77%
Frames mit $e_t \leq 20cm$	31,31%
Frames mit $e_t \leq 30cm$	56,82%
Frames mit $e_t \leq 40cm$	75,22%
Frames mit $e_t \leq 50cm$	82,91%
Frames mit $e_t > 50cm$	17,09%
Durchschnittlicher Fehler $\bar{e_t}$	43,81cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	24,33cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	138,31cm

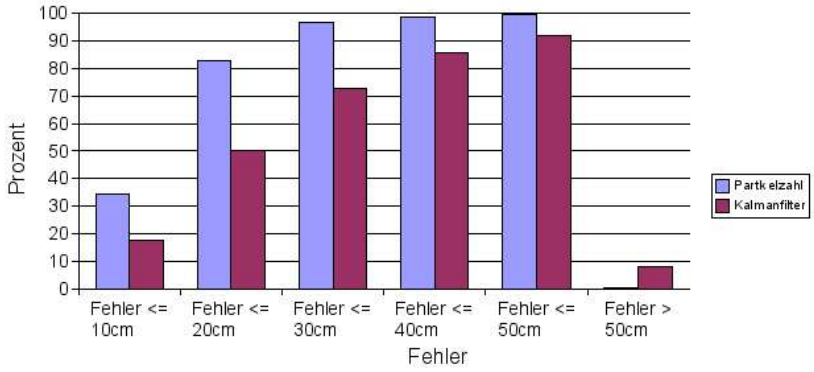

Abbildung 4.23: Trackingergebnisse auf der Sequenz „STZ“ von Partikelfilter und Kalmanfilter im Vergleich

4.7 Vergleich mit Kalmanfilter

Zum Abschluss der Experimente soll nun noch ein gewöhnlicher Kalmanfilter, wie in Abschnitt 2.4.1 vorgestellt, mit dem zuletzt verwendeten Partikelfilter verglichen werden. Der Kalmanfilter verwendet dabei die Einschränkung der Kamerazahl sowie die Verfolgung des Kopfes. Weitere Verbesserungen, die für Partikelfilter zusätzlich eingeführt wurden, sind auf Grund der Art des Kalmanfilter nicht möglich. Es ergeben sich für den Datensatz „STZ“ folgende Ergebnisse:

Tabelle 4.14: Fehlervergleich zwischen Partikelfilter und Kalmanfilter auf der Sequenz „STZ“

Bildfolge „STZ“		
Fehler	**Partikelfilter**	**Kalmanfilter**
Frames mit $e_t \leq 10cm$	34,24%	17,72%
Frames mit $e_t \leq 20cm$	82,98%	50,20%
Frames mit $e_t \leq 30cm$	96,72%	72,98%
Frames mit $e_t \leq 40cm$	98,74%	85,92%
Frames mit $e_t \leq 50cm$	99,79%	92,23%
Frames mit $e_t > 50cm$	0,21%	7,77%
Durchschnittlicher Fehler $\bar{e_t}$	13,78cm	27,54cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t \leq 50cm$	13,69cm	20,44cm
Durchschnittlicher Fehler $\bar{e_t}$ mit $e_t > 50cm$	56,14cm	111,98cm

Für den zweiten Datensatz „MYR“ verliert der Kalmanfilter die Spur, während der Sprecher zwischen Frame 4000 und 6000 an der Tafel steht, so dass ein Vergleich hier unzweckmäßig erscheint.

4.8 Übersicht

Abschließend soll nochmals grafisch veranschaulicht werden, wie sich das Trackingergebnis durch die Einführung verschiedener Modelle verbessert hat. Dabei zeigt sich, dass insbesondere die Einführung einer Mindestzahl an beobachtenden Kameras sowie das a-priori Aufenthaltsmodell die Ergebnisse deutlich verbessern konnten. Die Veränderung von Partikelzahl und Dynamikparametern hatte vor allem Einfluss auf die Genauigkeit der Schätzung.

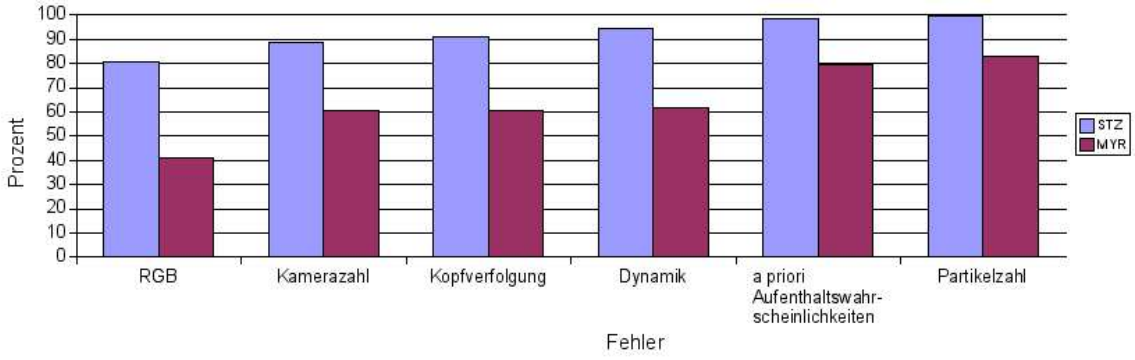

Abbildung 4.24: Entwicklung der Anzahl von Frames mit $e_t \leq 50cm$

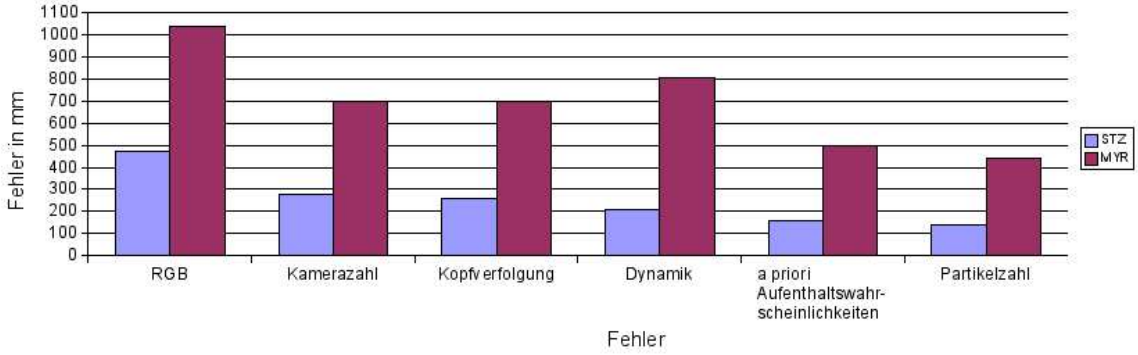

Abbildung 4.25: Entwicklung des durchschnittlichen Fehlers $\bar{e}_t$

5 Zusammenfassung und Ausblick

Das Verfolgen von Sprechern stellt eine wichtige Grundlage dar, um in intelligenten Räumen weitere Dienste anbieten zu können, wie zum Beispiel die Steuerung der Infrastruktur, in etwa Licht und Projektionsgeräte. In vorliegender Arbeit wurde das Verfolgen von Personen aufbauend auf adaptiven Hintergrundmodellen und Aufenthaltshypothesen im dreidimensionalen Raum mit Hilfe von Partikelfiltern im Vergleich zu Kalmanfiltern weiter verbessert. Dabei zeigte sich, dass der Einfluss von Zuschauerbewegungen durch den Einsatz eines a-priori Aufenthaltsmodells aus einer Mischung von Normalverteilungen sehr gut reduziert werden konnte. Voraussetzung allerdings war, dass sich die Zuschauer hinreichend weit vom Sprecher entfernt befanden. Außerdem zeigte sich, dass sich die Verfolgung des Sprechers insbesondere dann als schwierig gestaltet, wenn dieser zum Hintergrund adaptiert wurde oder sein Vordergrundbereich in mehrere kleine Teilbereiche zerfallen ist. Dies hatte weiterhin zur Folge, dass die Berechnungszeit für die 3D-Hypothesen stark anstieg. Zukünftige Arbeiten sollten sich deshalb mit folgenden möglichen Verbesserungsmöglichkeiten beschäftigen:

Segmentierung: Da insbesondere beim Seminarausschnitt „Mayer“ Probleme bei der Segmentierung auftraten, sollte überlegt werden, zum Algorithmus von Stauffer [17] eine Alternative zu finden. Das könnte zum ersten ein Segmentierungsalgorithmus sein, der auf der Erkennung von Farben basiert wie von King [12] bearbeitet. Darüber hinausgehend wäre der Einsatz kantenbasierter Verfahren in Erwägung zu ziehen, wie dies zum Beispiel bei Isard [10] geschieht. Ein solches wäre insbesondere robuster gegenüber stehenden Sprechern und weniger anfällig für Aktionen, die im Publikum stattfinden.

Laufzeit: Das vorgestellte System läuft nicht in Echtzeit, was insbesondere der Segmentierung und dem Triangulationsverfahren zu Lasten gelegt werden muss. Dieses wird nämlich umso laufzeitintensiver, je mehr extrahierte Vordergrundregionen existieren. Ist es bei zwei bis drei Regionen noch hinnehmbar zu versuchen, alle Teilmengen von möglichen Beobachtungen zu einer 3D-Hypothese zu kombinieren, verlängert es dennoch die Laufzeit exponentiell bei zerfallenen Vordergrundregionen des Sprechers. Eventuell wäre es hier möglich, die Segmentierungsergebnisse einer Kamera in das Bild der anderen zu projizieren, vergleichbar dem Epipolarebenenverfahren bei Stereokameras, um den Suchraum einzuschränken. Außerdem könnte man die letzte Schätzung in die Kamerabilder zurückprojizieren und nur die Vordergrundregionen in der Umgebung der Projektion zur Bildung der neuen

3D-Hypothesen heranziehen, was allerdings für nahe an der Kamera stehende Personen problematisch ist, wenn diese sich zu schnell bewegen und dies im Kamerabild aufgrund der perspektivischen Verzerrung einen großen Abstand ergibt.

Beobachtungsmodell: Eine weitere Möglichkeit, das aufwändige Triangulationsverfahren zu umgehen wäre es, die Partikelmenge auf der Basis der zweidimensionalen Kamerabilder zu bewerten. Dazu müssten die Ortskoordinaten der Statusvektoren in die vier Kamerabilder projiziert und anschließend mit einem geeigneten Beobachtungsmodell gewichtet werden. Dies wäre wesentlich schneller zu berechnen, da lediglich der Statusvektor mit der passenden Rotationsmatrix der Kamera multipliziert und anschließend der Translationsvektor addiert werden müsste, wodurch die aufwändige Lösung des überbestimmten Gleichungssystems entfallen würde. Dieser Ansatz wurde unter anderem bei Zotkin et. al. [21] mit Erfolg verwendet.

Verfolgung mehrer Personen: Die Verfolgung mehrer Personen würde gleichzeitig das Problem durchs Bild gehender Personen lösen. Einen entsprechenden Ansatz verfolgen Isard und MacCormik [11] mit dem System BraMBLe. Im vorgestellten System müsste dazu der Statusvektor um die Position und Geschwindigkeit weiterer Personen erweitert werden, was allerdings das Problem aufwirft, dass die genau Zahl der gerade zu beobachtenden Personen meist unbekannt ist.

Akustische Schätzergebnisse: Das Beobachtungsmodell könnte wie bei Zotkin [21] auch dadurch verbessert werden, dass neben einer visuellen Beobachtung die Ergebnisse einer akustischen Schätzung als weitere Modalität zur Neubewertung der Samplemenge mit einfließt. Dies könnte zumindestens zum Teil das Versagen der Segmentierung an manchen Stellen ausgleichen.

Bewegungsmodell: Auch am Bewegungsmodell wäre eine weitere Verbesserung möglich, da das hier benutzte lineare Bewegungsmodell einige Schwächen besitzt, insbesondere wenn der Sprecher spontan die Bewegungsrichtung wechselt. Hierzu wäre es denkbar, den Bewegungszustand des Sprechers eventuell mit einem Hidden Markov Modell mit Zuständen für unterschiedliche Bewegungsrichtungen und -geschwindigkeit und einem Zustand für „stehend“ zu verwenden, bei dem die Partikelmenge dann je nach Zustand unterschiedlich propagiert werden kann.

A-priori Aufenthaltsmodell: Das um a-priori Aufenthaltswahrscheinlichkeiten gewichtete Beobachtungsmodell könnte dahingehend verbessert werden, dass die Mittelwerte und Standardabweichungen nicht mehr von anderen Seminaren gelernt werden, sondern dass das Modell durch bisherige Schätzungen des gleichen Seminars mit Hilfe des EM-Algorithmus aufgebaut wird.

A Verwendete Parameterkonfigurationen

Die wichtigsten verwendeten Parameter in der Grundkonfiguration, die für den Farbraumvergleich RGB-YRG verwendet wurden, sollen im folgenden kurz aufgelistet werden. Die Variation der Parameter für die restlichen Experimente wird in Kapitel 4 beschrieben. Auf die Angabe der Einheiten wird hier auf Grund der Lesbarkeit verzichtet. Zur Einschätzung der Größenordnung sei allerdings angemerkt, dass Längen in mm gemessen werden und Zeiten in Anzahl von Bildframes.

Staufferalgorithmus zur Segmentierung

Normalverteilungen pro Pixel	1
σ für initiale Normalverteilungen	0,04
Lernrate α	0,00001
Anzahl erlaubter Standardabweichungen σ für Hintergrund	2,5
Größe des morphologischen Filters	3
Typ des morphologischen Filters	Schließend
Mindestgröße in Pixeln für segmentierte Regionen	25

Triangulation

Maximal erlaubtes Residuum	200
Mindestzahl von Kameras die Sprecher beobachten	2

Partikelfilter

Anfängliche Partikelverteilung μ	$\mu = \begin{pmatrix} 500 \\ 3500 \\ 1000 \end{pmatrix}$
Anfängliche Partikelverteilung Σ	$\Sigma = \begin{pmatrix} 500^2 & 0 & 0 \\ 0 & 500^2 & 0 \\ 0 & 0 & 500^2 \end{pmatrix}$
Anzahl verwendeter Partikel	500
μ_d für Bewegungsmodell	$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$
Σ_d für Bewegungsmodell	$\begin{pmatrix} 4^2 & 0 & 0 \\ 0 & 4^2 & 0 \\ 0 & 0 & 2^2 \end{pmatrix}$
σ_o für Beobachtungsmodell	$\sigma = 300$

Kalman Filter

Anfängliche Auftrittsverteilung	$\mu = \begin{pmatrix} 500 \\ 3000 \\ 1000 \end{pmatrix}$
Q	$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 16 & 0 & 0 \\ 0 & 0 & 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}$
R	$\begin{pmatrix} 90000 & 0 & 0 \\ 0 & 90000 & 0 \\ 0 & 0 & 90000 \end{pmatrix}$

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Bouguet, Jean-Yves: *Camera Calibration Toolbox for Matlab, Verfügbar unter:* `http://www.vision.caltech.edu/bouguetj/calib_doc`, 2001.

[2] Cai, Q. und J. Aggarwal: *Tracking Human Motion Using Multiple Cameras*, 1996.

[3] Checka, Neal, Kevin Wilson, Michael Siracusa und Trevor Darrell: *Multiple Person and Speaker Activity Tracking with a Particle Filter.* In: *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[4] Darrell, T., G. Gordon, M. Harville und J. Woodfill: *Integrated person tracking using stereo, color, and pattern detection.* International Journal of Computer Vision, 37(2):175–185, 2000.

[5] Focken, Dirk: *Adaptive Hintergrundmodelle zur Personenverfolgung.* Studienarbeit, Universität Karlsruhe, 2001.

[6] Focken, Dirk: *Vision-based 3-D Tracking of People in a Smart Room Environment.* Diplomarbeit, Universität Karlsruhe, 2002.

[7] Focken, Dirk und Rainer Stiefelhagen: *Towards Vision-based 3-D People Tracking in a Smart Room.* In: *Proceedings of the 2002 International Conference on Multimodal Interfaces (ICMI '02)*, October 2002.

[8] Grenander, Ulf, Yun-Shyong Chow und D. M Keenan: *Hands: A Pattern Theoretic Study of Biological Shapes.* Springer-Verlag New York Inc., 1991.

[9] Haritaoglu, I., D. Harwood und L. Davis: *W4: Who? when? where? what? a real time system for detecting and tracking people.* Face and Gesture Recognition, Seiten 222–227, 1998.

[10] Isard, Michael und Andrew Blake: *Condensation – conditional density propagation for visual tracking.* International Journal of Computer Vision, 29(1):5–28, 1998.

[11] Isard, Michael und John MacCormik: *BraMBLe: A Bayesian Multiple-Blob Tracker*. In: *Proceedings of International Conference on Computer Vision*, Band 2, Seiten 34–41, 2001.

[12] King, Andy: *Farbbasierte Segmentierung von Körperregionen*. Studienarbeit, Universität Karlsruhe, 2002.

[13] Krumm, John, Steve Harris, Brian Meyers, Barry Brumitt, Michael Hale und Steve Shafer: *Multi-Camera Multi-Person Tracking for Easy Living*. In: *Third IEEE International Workshop on Visual Surveillance, July 1*, 1997.

[14] Mikic, I., S. Santini und R. Jain: *Tracking Objects in 3D using Multiple Camera Views*. In: *Proceedings of Asian Conference on Computer Vision*, 2000.

[15] Nickel, Kai: *Erkennung von Zeigegesten basierend auf 3D-Tracking von Kopf und Händen*. Diplomarbeit, Universität Karlsruhe, 2003.

[16] Russell, Stuart und Peter Norvig: *Artificial Intelligence – A Modern Approach*, Seiten 551–559. Prentice Hall, Upper Saddle River, New Jersey 07458, Zweite Auflage, 2003.

[17] Stauffer, Chris und Eric Grimson: *Adaptive background mixture models for realtime tracking*. In: *Proceedings of CVPR*, Seiten 333–339, 1998.

[18] Ward, D.B. und R.C. Williamson: *Particle filter beamforming for acoustic source localization in a reverberant environment*. In: *Proceedings of IEEE ICASSP*, Band 2, Seiten 1777–1780, Mai 2002.

[19] Wren, C., A. Azarbayejani, T. Darrell und A. Pentland: *Pfinder: Real-time tracking of the human body*. In: *Photonics East, SPIE, volume 2615, 1995. Bellingham, WA*, 1999.

[20] Zhang, Zhengyou: *Flexible Camera Calibration by Viewing a Plane from Unknown Orientations*. In: *Proceedings of International Conference on Computer Vision*, Seiten 666–673, 1999.

[21] Zotkin, Dmitry N., Ramani Duraiswami und Larry S. Davis: *Joint Audio-Visual Tracking using Particle Filters*. EURASIP Journal on Applied Signal Processing Special Issue on Joint Audio-Visual Speech Processing, 2002.