in: Proc. 3rd Workshop on Self-Organization of AdaptiVE Behavior (SOAVE), 2004, Ilmenau, pp. 27-36
Fortschritt-Berichte VDI, Reihe 10, Nr. 743, VDI-Verlag 2004

Vergleich von hautfarbbasierten Multi-Target-Trackern

T. Wilhelm und C. Martin
Technische Universität Ilmenau, Fachgebiet Neuroinformatik
PF 100565, 98684 Ilmenau
{Torsten.Wilhelm, Christian.Martin}@TU-Ilmenau.de

Abstract

Für den Einsatz auf einem mobilen Robotersystem wurden zwei hautfarbbasierte Multi-Target-Tracker implementiert. Um die Hautfarberkennung unabhängig von variierenden Beleuchtungsverhältnissen zu machen, wurde ein automatischer Weißabgleich entwickelt. Beide Verfahren sind Erweiterungen des klassischen Partikelfilters, wobei eines eine hochdimensionale Sample-Konfiguration verwendet, um mehrere Objekte zu beschreiben und das andere mehrere einzelne Partikelfilter. Beide Verfahren werden theoretisch hinsichtlich der verwendeten Heuristiken und praktisch anhand von experimentellen Untersuchungen gegenübergestellt.

1 Einleitung

Der Einsatz von Servicerobotern in realen Umgebungen gewinnt in den letzten Jahren immer mehr an Bedeutung. Die meisten dieser Roboter stellen allerdings eine bestimmte Dienstleistung bereit, ohne dabei aktiv mit einem Nutzer kommunizieren zu können. Ein Grund hierfür mag die Schwierigkeit der robusten visuellen Detektion von Personen unter realen Einsatzbedingungen sein. Trotzdem ist es für eine natürliche Mensch-Maschine-Kommunikation unerlässlich, Personen auch auf mobilen Systemen während der Ausführung einer Serviceleistung robust zu detektieren und zu verfolgen. Falls der Nutzer unaufmerksam ist, sich anderen Dingen zuwendet oder die Kommunikation ganz abbricht, sollte der Roboter geeignete Aktionen ausführen und nicht stupide weiter seine Serviceaufgabe ausführen. Beispielhaft sei hier unser Serviceroboter PERSES genannt, der für den Einsatz in einem Baumarkt entwickelt wird und seine Nutzer zu von ihnen gewünschten Produkten lotsen soll [8]. Dabei werden potentielle Nutzer aufgrund von Hautfarbe in einem Panoramabild detektiert und robust verfolgt. Da zu jedem Zeitpunkt mehrere Hautfarbregionen im Bild vorhanden sein können und die Zuordnung zu Nutzer, andere Person oder falsch-positive nicht frühzeitig getroffen werden kann, muss das Tracking-System in der Lage sein, mehrere Hautfarbregionen gleichzeitig zu tracken. Für die Lösung dieses Problems wurden zwei alternative Ansätze zum Multi-Target-Tracking implementiert und verglichen. In Abschnitt 2 wird auf die Probleme mit dem für das Tracking verwendeten Feature Hautfarbe eingegangen. Abschnitt 3 erläutert den CONDENSATION-Algorithmus als Grundlage der beiden Tracker. Die Problematik des Multi-Target-Trackings und die beiden implementierten Verfahren werden im Abschnitt 4 behandelt. Kapitel 5 stellt die vergleichenden Untersuchungen und die erzielten Ergebnisse vor.

2 Hautfarbe

Hautfarbe ist ein häufig verwendeter Cue bei der Suche nach Personen in Bildern. Sie bietet den Vorteil, unabhängig von Eigenbewegungen des Roboters berechnet werden zu können und eignet sich dadurch besonders für den Einsatz auf mobilen Systemen. Allerdings funktioniert dies nur dann zufriedenstellend, wenn die Beleuchtungsverhältnisse bei der Aufnahme der Trainingsdaten und bei der Anwendung hinreichend ähnlich sind, was in der Regel nicht gewährleistet werden kann. Ein entscheidendes Problem bei der hautfarbbasierten Nutzerdetektion sind die stark veränderlichen Beleuchtungsbedingungen, mit denen bei mobilen Systemen zwangsläufig gerechnet werden muss. In typischen Realwelt-Einsatzfeldern können die Beleuchtungsverhältnisse von reinem Kunstlicht bis zu reiner natürlicher Beleuchtung variieren.

Farbraum Für unsere Untersuchungen verwenden wir den helligkeitsnormierten dichromatischen r-g-Farbraum [4], da er sehr gut geeignet ist, Hautfarbe in einem großen Bereich unterschiedlicher Beleuchtungsverhältnisse zu repräsentieren [9]. Die Farbwerte in diesem Raum ergeben sich aus $r = R/(R+G+B)$ und $g = G/(R+G+B)$. Im Prinzip ist aber jeder Farbraum geeignet, bei dem Helligkeits- und Farbinformationen dekorreliert vorliegen.

Farbmodell Das verwendete Farbmodell besteht aus einer Lookup-Tabelle mit manuell als Hautfarbe klassifizierten Pixeln im r-g-Farbraum. Ein ähnlicher Ansatz, allerdings im YUV-Farbraum, wurde in [3] vorgestellt. Das verwendete Farbmodell ist in Abbildung 1(a) zu sehen. Die Varianzen dieser Verteilung hängen zum einen von den Variationen der Hautfarben der verschiedenen Probanden ab, aber auch von den Beleuchtungsschwankungen während der Aufnahme der Trainingsdaten. Nur wenn letztere möglichst gering gehalten werden, kann ein spezifisches Hautfarbmodell erzeugt werden. Verglichen mit der Hautfarbverteilung in [1] hat die Hautfarbverteilung in Abbildung 1(a) eine ähnliche Form (Skin locus) [5], aber eine wesentlich geringere Ausdehnung. Dies ist darauf zurückzuführen, dass die Daten für dieses Modell unter Verwendung des im nächsten Abschnitt beschriebenen automatischen Weißabgleichs aufgenommen wurden.

Automatischer Weißabgleich Um mit dem Problem der variierenden Beleuchtungsbedingungen umgehen zu können, wurde ein automatischer Weißabgleich für die omnidirektionale Kamera entwickelt. Für diesen Zweck wurde die Kamera mit einem weißen Aluminumring ausgestattet, der als Weißreferenz dient. Abb. 1(b) zeigt die Kamera mit omnidirektionalem Speigel und Weiß-Referenz, und Abb. 1(c) zeigt ein mit dieser Kamera aufgenommenes Bild, bei dem sich die Weiß-Referenz auf einem innen gelegenen Radius befindet. Die Oberfläche des Rings ist nicht horizontal und flach, sondern weist eine leichte konvexe Krümmung auf, so das auch seitlich einfallendes Licht berücksichtigt wird. Der automatische Weißabgleich verwendet die Möglichkeit der Digitalkamera (SONY DFW VL500), Parameter für den Weißabgleich für U und V (YUV-Farbraum) einzustellen. Es werden die Mittelwerte für R, G und B über alle Pixel, die innerhalb der Weißreferenz liegen, berechnet und in den YUV-Farbraum transformiert. Die Abweichungen von U und V von den Sollwerten $U = 0$ und $V = 0$ dienen als Eingangsgrößen für zwei separate diskrete PID-Regler. Diese berechnen die entsprechenden Parameter für den Weißabgleich der Kamera [7]. Außerdem kann der mittlere Wert für Y für die Regelung der Iris der Digitalkamera verwendet werden, so dass auch eine relativ konstante Helligkeit erreicht wird.

Abbildung 2 zeigt die Auswirkungen des automatischen Weißabgleiches auf die Hautfarbdetektion. Die Stellgrößen für den Weißabgleich sind direkt nach dem Einschalten des Systems mit Vorgabewerten belegt. Nach nur zehn Bildern (rund 1 Sekunde) ist die Regelabweichung nahezu null.

(a) (b) (c)

Abbildung 1. Automatischer Weißabgleich auf Bildern der omnidirektionalen Kamera. (a) Lookup-Tabelle für Hautfarbe im dichromatischen r-g-Farbraum. (b) Weiß-Referenz zwischen Kamera und Objektiv. (c) Mit dieser Kamera aufgenommenes Bild, wobei die Weiß-Referenz in der Nähe des Bildzentrums erscheint (markiert mit einem Karo-Muster). Diese Bildregion ist für die Nutzerdetektion und das Nutzertracking nicht von Interesse, da hier der Fußboden in unmittelbarer Umgebung des Roboters abgebildet wird.

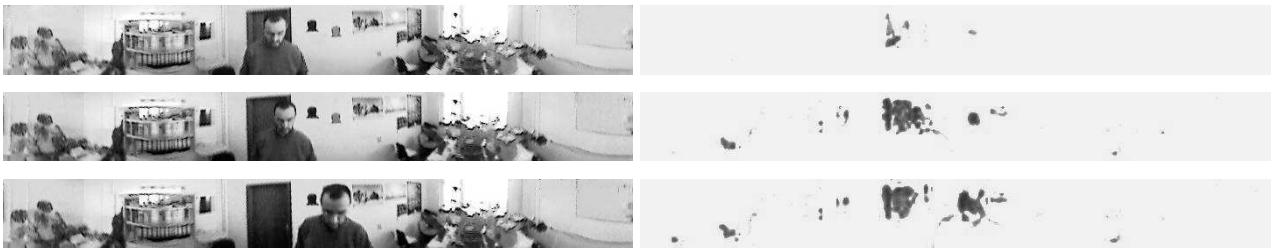


Abbildung 2. Zunehmende Hautfarbdetektionen durch den automatischen Weißabgleich. Es handelt sich um das erste, das fünfte und das zehnte Bild einer Sequenz. Neben der eigentlichen Hautfarbregion entstehen auch mehr falsch-positive Detektionen auf holzfarbenen Objekten.

3 Condensation-Algorithmus

Die Basis der beiden untersuchten alternativen Tracking-Systeme bildet der CONDENSATION-Algorithmus [2]. Die Aufgabe der Berechnung der Wahrscheinlichkeit, ob sich an einem bestimmten Bildpunkt ein Gesicht/Hautfarbe befindet und die Verfolgung der resultierenden Dichtefunktion über die Zeit t, wird durch eine Approximation der Dichtefunktion $p(\mathbf{x}_t)$ durch eine relativ kleine Anzahl von N Samples $\mathbf{s}_t^{(i)}$ realisiert:

$$p(\mathbf{x}_t) \propto \left\{ \mathbf{s}_t^{(i)} = \left\langle \mathbf{x}_t^{(i)}, w_t^{(i)} \right\rangle | i = 1, .., N \right\} \tag{1}$$

wobei jedes Sample $\mathbf{s}_t^{(i)}$ durch eine Position $\mathbf{x}_t^{(i)}$ und ein Gewicht $w_t^{(i)}$ charakterisiert wird. Laut [2] läuft ein Aktualisierungsschritt des rekursiven Filters wie folgt ab:

$$P(\mathbf{s}_t|\mathbf{Y}_t) = \underbrace{P(\mathbf{y}_t|\mathbf{s}_t)}_{\text{Beobachtungsmodell}} \int \underbrace{P(\mathbf{s}_t|\mathbf{s}_{t-1})}_{\text{Bewegungsmodell}} \cdot P(\mathbf{s}_{t-1}|\mathbf{Y}_{t-1}) d\mathbf{s}_{t-1} \tag{2}$$

Wir beginnen mit einer Menge von Samples $\mathbf{s}$, welche die a posteriori Dichte $p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1})$ aus dem vorherigen Zeitschritt beschreibt, wobei $\mathbf{Y}_{t-1}$ die Menge aller bisherigen Beobachtungen $\{\mathbf{y}_0, ..., \mathbf{y}_{t-1}\}$

ist. $\mathbf{s}$ wird entsprechend des Bewegungsmodells $P(\mathbf{s}_t|\mathbf{s}_{t-1})$ propagiert, welches aus einer stochastischen Komponente für unverhergesehene Bewegungen der Person und aus einer deterministischen Komponente für bekannte Bewegungen des Roboters besteht. Man erhält so die neue Sample-Menge $\mathbf{s}'$, die die a priori Dichte $p(\mathbf{x}_t|\mathbf{Y}_{t-1})$ beschreibt. Der nächste Schritt besteht in einem faktorisierten Resampling, bei dem die neuen Sample-Gewichte $w_t^{(i)}$ entsprechend der Beobachtungen vom Hautfarbdetektor im aktuellen Zeitschritt zugewiesen werden. Daraufhin werden Samples mit der Wahrscheinlichkeit $w_t^{(i)}$ aus der Menge $\mathbf{s}'$ ausgewählt. Die so entstandene Sample-Menge $\mathbf{s}''$ repräsentiert die a posteriori Dichtefunktion $p(\mathbf{x}_t|\mathbf{Y}_t)$.

Im Vergleich zu einem Panoramabild mit 720×106 Pixeln wird mit dem Condensation-Algorithmus die Feature-Extraktion für nur 500 Samples berechnet, was einer Reduktion auf 0.655% entspricht. Das Zentrum der resultierenden Sample-Verteilung dient als Hypothese für die Position eines Gesichtes.

4 Multi-Target-Tracking

Obwohl es für den Serviceroboter eigentlich immer nur ein interessantes und damit zu trackendes Objekt gibt, nämlich seinen aktuellen Nutzer, sollte das Tracking-System in der Lage sein, mehrere Objekte gleichzeitig zu verfolgen. Meldet sich der aktuelle Nutzer ab und verlässt den Einsatzbereich des Roboters, kann dieser sich so unmittelbar einer anderen Person zuwenden. Da das Tracking-System lediglich über Hautfarbinformationen verfügt, kann es keine Aussage darüber treffen, ob es sich um ein Gesicht oder ein anderes hautfarbenes Objekt handelt, so dass zunächst alle Hautfarbregionen als potentielle Nutzer in Betracht gezogen werden müssen. Mit einem nachgeschalteten Gesichtsdetektor ist es dann möglich, einzelne fehlerhafte Hautfarbregionen als falsch-positiv zu markieren. Auch in diesem Fall sollten diese weiter verfolgt werden, so dass sie dem System als Falsch-positive bekannt bleiben.

Theoretisch ist ein einzelner Condensation-Tracker in der Lage, mehrere Objekte gleichzeitig zu verfolgen, da er jede beliebige multimodale Dichtefunktion approximieren kann. In der Regel konzentriert sich ein einzelner Tracker aber auf den stärksten Stimulus, da hier die meisten Samples neu entstehen und die Anzahl der Samples auf anderen Regionen immer kleiner wird. Um trotzdem mehrere Stimuli tracken zu können, wurden zwei Ansätze implementiert und vergleichend untersucht, die in den folgenden Abschnitten beschrieben werden.

4.1 Verwendung von hochdimensionalen Sample-Konfigurationen

In diesem Abschnitt wird ein Verfahren zum Multi-Objekt-Tracking basierend auf der Arbeit von Tao et al. [6] vorgestellt. Hier wird eine sogenannte Sample-Konfiguration eingeführt, die die Kombination der Zustände aller beobachteten Objekte in einem Bild beschreibt. Eine solche Konfiguation kann wie folgt ausgedrückt werden:

$$\mathbf{c}_t^{(i)} = \left\langle \left\{ \mathbf{x}_{t,1}^{(i)}, \mathbf{x}_{t,2}^{(i)}, ..., \mathbf{x}_{t,m}^{(i)} \right\}, w_t^{(i)} \right\rangle | i = 1, .., N; j = 1, .., M \tag{3}$$

wobei $\mathbf{x}_{t,j}$ der Zustand des Objektes j zum Zeitpunkt t (z.B. die Koordinaten einer Person im Bild) und M die Anzahl der Samples in der Konfiguration ist. Eine solche hochdimensionale Sample-Konfiguration $\mathbf{c}_t$ beschreibt also den Zustand von M Objekten in einer einzigen Variablen. Ziel des Verfahrens ist es, die a posteriori Wahrscheinlichkeit der Konfigurations-Parameter mit Hilfe eines geeigneten Bayes-Filters zu bestimmen:

$$P(\mathbf{c}_t|\mathbf{Y}_t) = \underbrace{P(\mathbf{y}_t|\mathbf{c}_t)}_{\text{Konfigurationsgüte}} \int \underbrace{P(\mathbf{c}_t|\mathbf{c}_{t-1})}_{\text{Konfigurationsdynamik}} \cdot P(\mathbf{c}_{t-1}|\mathbf{Y}_{t-1}) d\mathbf{c}_{t-1} \tag{4}$$

Der Term $P(\mathbf{y}_t|\mathbf{c}_t)$ stellt die Konfigurationsgüte dar und ist ein Maß dafür, wie gut die aktuelle Beobachtung $\mathbf{y}_t$ durch die Konfiguration $\mathbf{c}_t$ beschrieben werden kann. Dieser Term entspricht dem Beobachtungsmodell im klassischen CONDENSATION-Algorithmus, vgl. Gleichung 2. Die Konfigurationsdynamik $P(\mathbf{c}_t|\mathbf{c}_{t-1})$ beschreibt, wie sich eine Konfiguration $\mathbf{c}_{t-1}$ zur Konfiguration $\mathbf{c}_t$ verändert. Im klassischen CONDENSATION-Algorithmus ist dies das Bewegungsmodell. In den beiden folgenden Abschnitten werden diese beiden Terme ausführlicher beschrieben.

4.1.1 Konfigurationsgüte

Die Konfigurationsgüte $P(\mathbf{y}_t|\mathbf{c}_t)$ ist eine sehr komplexe und unter Umständen auch sehr schwierig zu berechnende Verteilung. In unserer Arbeit haben wir daher die gleiche Dekomposition der Konfiguration wie in [6] eingesetzt. Tao et al. haben dazu eine Energiefunktion definiert, die erwünschten Konfigurationen hohe Werte und nicht erwünschten Konfigurationen niedrige Werte zuweist. Die Funktion besteht aus drei Faktoren:

1. **Objektwahrscheinlichkeit** λ: Dieser Faktor gibt an, wie gut die Objekte der Konfiguration $\mathbf{c}_t$ mit Hilfe der aktuellen Beobachtung $\mathbf{y}_t$ erklärt werden können. Dazu werden die Wahrscheinlichkeiten der einzelnen Samples $\mathbf{x}_{t,j}$ der Sample-Konfiguration genutzt. Die Wahrscheinlichkeit $H(\mathbf{x}_{t,j})$ ist die Messung des Hautfarbdetektors an der Position des Samples $\mathbf{x}_{t,j}$. Die Verwendung des geometrischen Mittels bewirkt, dass Konfigurationen, bei denen nur ein Sample nicht auf einer Hautfarbregion liegt, bereits sehr schlecht bewertet werden.

$$\lambda(\mathbf{c}_t) = \left(\prod_{j=1}^{m} H(\mathbf{x}_{t,j}) \right)^{\frac{1}{m}} \tag{5}$$

2. **Konfigurationsabdeckung** γ: Dieser Term gibt an, wie gut eine Konfiguration die gesamte Beobachtung abdeckt. Er wird wie folgt berechnet:

$$\gamma(\mathbf{c}_t) = \frac{\left| A \cap \left(\bigcup_{j=1}^{m} B_j \right) \right| + b}{|A| + b} \tag{6}$$

   wobei A die Menge aller Einzelbeobachtungen, d.h. separater Hautfarbregionen, und B_j der Zustand des Objektes j der Konfiguration ist. Die Schnittmenge $A \cap \left(\bigcup_{j=1}^{m} B_j \right)$ ist somit die Menge aller Beobachtungen, die auch durch die zu bewertende Konfiguration abgedeckt werden. Der Faktor γ geht somit dann gegen 1.0, wenn alle Beobachtungen auch durch die entsprechende Konfiguration erfasst werden. Wenn dagegen keine einzige Beobachtung erfasst wird, geht γ gegen 0.0. Die (kleine) positive Konstante b verhindert eine Division durch 0.

   Um diese Größe berechnen zu können, muss allerdings die Anzahl von Objekten in der Szene bereits bekannt sein. Um diese zu schätzen, wird das Bild stark unterabgetastet und für jeden Pixel dieses unterabgetasteten Bildes die Hautfarbzugehörigkeit mit einem Schwellwert ermittelt. Die Summe dieser Zugehörigkeiten ist eine Schätzung für die Größe A.

3. **Konfigurationskompaktheit** ξ: Die Kompaktheit ist definiert als das Verhältnis der repräsentierten Beobachtungen zur Komplexität der Konfiguration $\mathbf{c}_t$. Sie wird wie folgt berechnet:

$$\xi(\mathbf{c}_t) = \frac{\left|A \cap \left(\bigcup_{j=1}^{m} B_j\right)\right| + d}{\left|\bigcup_{j=1}^{m} B_j\right| + a} \tag{7}$$

Dieser Wert geht dann gegen 1.0, wenn eine vollständige Abdeckung der Daten erfolgt. Wenn in einer Konfiguration zu wenige oder zu viele Samples verwendet werden, um eine bestimmte Menge von Einzelbeobachtungen zu repräsentieren, wird ξ klein. d ist eine positive Konstante, die so gewählt wird, dass wenn $|A| = 0$, Konfigurationen mit weniger Samples eine höhere Bewertung bekommen. Der (kleine) positive Wert a verhindert eine Division durch 0.

Konfigurationsgüte: Letztendlich wird die Güte einer Konfiguration $\mathbf{c}_t$ approximiert durch:

$$P(\mathbf{y}_t|\mathbf{c}_t) \approx \lambda(\mathbf{c}_t) \cdot (\gamma(\mathbf{c}_t) \cdot \xi(\mathbf{c}_t))^{\delta} \tag{8}$$

wobei δ eine positive Konstante ist, die die relative Wichtigkeit der letzten beiden Faktoren beeinflusst. Die Werte $P(\mathbf{y}_t|\mathbf{c}_t)$ werden nachträglich normiert und dienen im nachfolgenden Zeitschritt als Gewichte w_t für die Sample-Konfigurationen im Resampling-Schritt.

4.1.2 Konfigurationsdynamik

Das Bewegungsmodell entspricht dem des normalen CONDENSATION-Algorithmus und besteht aus einer stochastischen und einer deterministischen Komponente. Für das Einfügen bzw. Löschen von Objekten wird das Bewegungsmodell um zwei Wahrscheinlichkeiten erweitert. Mit der Wahrscheinlichkeit α wird ein neues Sample in eine Konfiguration eingefügt, wobei dessen Position zufällig initialisiert wird. Mit der Wahrscheinlichkeit β wird ein Sample aus einer Konfiguration gelöscht. In unserer Implementierung werden jeweils konstante Werte verwendet (z.B. $\alpha = 0.01$ und $\beta = 0.01$). Dieses erweiterte Bewegungsmodell wird als Konfigurationsdynamik bezeichnet, siehe Gleichung 4.

4.1.3 Schätzung der Objektanzahl und -positionen

Die geschätzte Anzahl von Objekten zum Zeitpunkt t kann wie folgt berechnet werden:

$$\sum_{i=1}^{N} \left|\mathbf{c}_t^{(i)}\right| w_t^{(i)} \quad \text{mit} \quad 0 \leq \left|\mathbf{c}_t^{(i)}\right| \leq M \tag{9}$$

wobei $\left|\mathbf{c}_t^{(i)}\right| \in \mathbb{N}$ die Anzahl der Samples $\mathbf{x}_{t,j}^{(i)}$ der Konfiguration $\mathbf{c}_t^{(i)}$ ist. Die Position der einzelnen Objekte im Bild wird wie folgt geschätzt: Da die Samples in den Konfigurationen nicht nach ihrer Objektzugehörigkeit geordnet vorliegen, wird versucht, diese entsprechend ihrer räumlichen Lage einander zuzuordnen. Hierzu wird der Abstand θ_d definiert, den zwei zu einem Objekt gehörende Samples nicht überschreiten dürfen.

4.1.4 Komplexität

Wenn in einer Szene M Objekte vorhanden sind, dann muss die a posteriori Verteilung im Raum X^M gesampelt werden. Um die gleiche Sample-Dichte beizubehalten, muss die Anzahl der Samples exponentiell bezüglich M sein, was diesen Algorithmus nur für kleine M praktikabel macht. Für große M (etwa $M > 10$) ist der beschriebene Algorithmus nicht mehr sinnvoll einsetzbar. Zur Lösung dieses Problems für große M wird in [6] ein hierarchischer Sampling-Algorithmus vorgestellt. Im Rahmen unserer Arbeiten wurde ein Wert von $M = 5$ verwendet, bei dem das oben beschriebene Verfahren noch direkt eingesetzt werden kann.

4.2 Verwendung von mehreren Einzel-Trackern

Das alternative selbst entwickelte Verfahren verwendet mehrere voneinander unabhängige CONDENSATION-Tracker, von denen jeder ein Objekt in der Szene verfolgt. Die Anzahl der Samples pro Tracker und die maximale Anzahl der verwendeten Tracker ist dabei prinzipiell beliebig, wobei die Fragen zu klären sind, wann ein Tracker eingefügt und wann ein Tracker gelöscht werden muss. Alle aktuell verwendeten Tracker werden bei diesem Verfahren in einer Liste verwaltet, wobei der erste Tracker in der Liste der aktuellen Nutzerhypothese entspricht. Die Reihenfolge in der Liste entspricht dann einer Priorität der Tracker.

Einfügen eines neuen Trackers Befindet sich ein Objekt im Bild, dass noch nicht von einem Tracker verfolgt wird, wird an der entsprechenden Bildstelle ein neuer Tracker eingefügt. Hierzu werden die Pixel des Bildes in einem diskreten Raster auf das Vorhandensein von Hautfarbe untersucht. Auf eine Unterabtastung des Bildes wie beim Verfahren von Tao wurde aus Rechenzeitgründen verzichtet. Liegt die mittlere Hautfarbzugehörigkeit für ein Rasterelement über einem Schwellwert θ_i und ist in unmittelbarer Nähe dieses Elements noch kein Tracker vorhanden, wird dort ein neuer Tracker initialisiert. Dieses Vorgehen entspricht der Berechnung der Konfigurationsabdeckung γ nach Tao. Allerdings wird hier nach der Detektion eines nicht-getrackten Hautfarbbereichs zielstrebig an genau dieser Stelle ein Tracker plaziert, während bei Tao Konfigurationen, die einen solchen Hautfarbbereich nicht erfassen, schlechter bewertet werden. Ein neuer Hautfarbbereich wird bei Tao erst dann erfasst, wenn in einer Konfiguration mit der Wahrscheinlichkeit α zufällig ein neues Sample an der entsprechenden Position im Bild erzeugt wird.

Löschen von vorhandenen Trackern Ein Tracker wird in den folgenden Fällen aus der Liste gelöscht:

1. Die mittlere Hautfarbwahrscheinlichkeit, d.h. das arithmetische Mittel der Messungen des Hautfarbdetektors an den Samplepositionen $\mathbf{x}_t^{(i)}$ unterschreitet einen Mindestwert θ_e. Dieser Fall tritt z.B. dann ein, wenn sich eine getrackte Person aus dem Umfeld des Roboters entfernt und die Hautfarbregion entsprechend kleiner wird. Dieses Maß entspricht der Objektwahrscheinlichkeit λ bei Tao, wobei hier für die Berechnung das arithmetische Mittel verwendet wird, da einzelne abweichende Samples das Gesamtergebnis nicht zu sehr beeinflussen sollen.

$$\lambda_j = \frac{1}{n} \sum_{i=1}^{n} H(\mathbf{x}_{t,j}^{(i)}) \tag{10}$$

2. Der Mindestabstand θ_d zwischen den Schwerpunkten zweier Tracker wird unterschritten. Damit ein Objekt nicht von mehreren Trackern verfolgt wird, sobald sich zwei Tracker zu nahe kommen, wird derjenige, der in der Liste weiter hinten steht, gelöscht. So wird sichergestellt, daß die aktuelle Nutzerhypothese, die an oberster Stelle in der Liste steht, nicht durch andere Tracker verdrängt werden kann. Dieser Fall wird bei Tao nur indirekt über die Konfigurationsabdeckung γ berücksichtigt. Bei der Unterabtastung des Bildes und der Schätzung der Anzahl der Objekte, werden zwei sehr nahe beeinanderliegende Objekte als eines gezählt. In diesem Fall werden Konfigurationen mit mehr Samples schlechter bewertet und sterben nach kurzer Zeit aus. Der Mindestabstand θ_d korrespondiert also mit dem Unterabtatstungsfaktor bei Tao.

4.2.1 Schätzung der Objektpositionen

Die geschätzten Objektpositionen $\mathbf{x}_j$ entsprechen bei diesem Verfahren dem Schwerpunkt der Sample-Verteilungen der einzelnen CONDENSATION-Tracker j. Diese werden wie folgt berechnet:

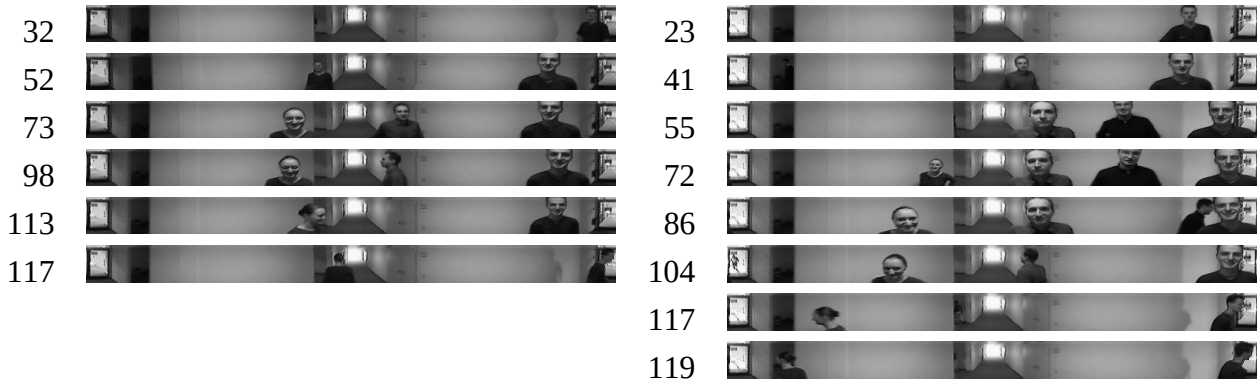


Abbildung 3. Zum Test verwendete Datensätze. (a) Datensatz 1 mit 3 Personen. (b) Datensatz 2 mit 4 Personen. Jeweils links neben einem Bild ist dessen Frame-Nummer im Datensatz angegeben. Die gezeigten Bilder sind genau die, bei denen eine Person die Szene betritt oder verlässt.

$$\mathbf{x}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_{t,j}^{(i)} w_{t,j}^{(i)} \tag{11}$$

5 Vergleichende Untersuchungen

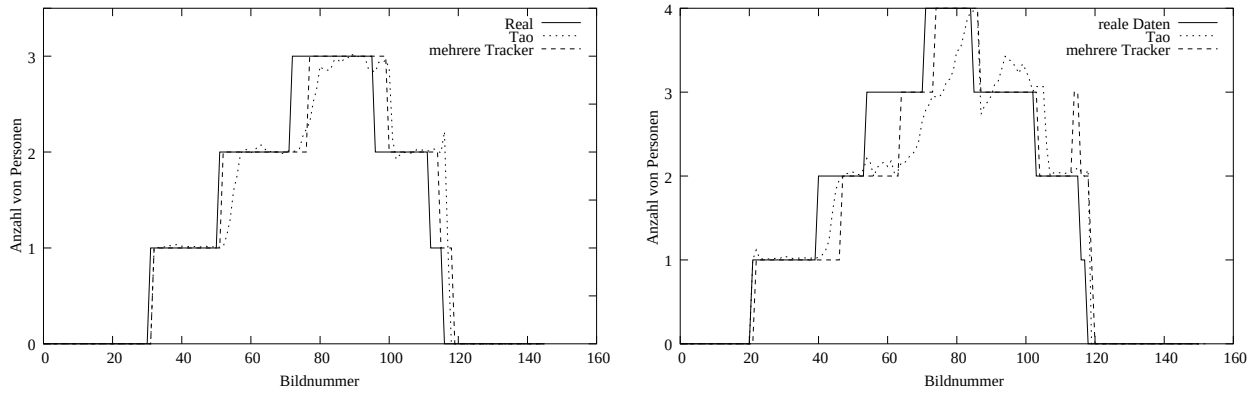
5.1 Datensatz

Abbildung 3 zeigt die beiden für den Vergleich verwendeten Datensätze. Beide Sequenzen beginnen mit einer leeren Szene, wobei nacheinander jeweils eine Person hinzukommt und zum Schluss alle Personen wieder verschwinden. Es sollte hierbei untersucht werden, inwieweit die beiden Systeme in der Lage sind, zu jedem Zeitpunkt die Anzahl der Personen in der Szene korrekt zu schätzen. Dazu wurde die Anzahl der vorhandenen Personen für jedes Bild von Hand ermittelt.

5.2 Ergebnisse

Abbildung 4 zeigt die Ergebnisse der beiden Tracking-Systeme auf den oben vorgestellten Datensätzen. Prinzipiell sind beide Tracker in der Lage, alle in den Sequenzen auftauchenden Personen zu erfassen. Das selbst entwickelte System, das mehrere Einzel-Tracker verwendet, ist in einigen Situationen zuverlässiger, da es die korrekte Anzahl von Personen schneller ermitteln kann.

Abbildung 5 zeigt den Zeitbedarf der beiden Tracking-Systeme auf den oben vorgestellten Datensätzen. Für beide Systeme wächst die benötigte Rechenzeit mit der aktuellen Anzahl der getrackten Objekte. Die Tests wurden auf einem Rechner mit einem AMD Athlon 2800+ durchgeführt. Beide Verfahren können als echtzeitfähig eingestuft werden und sind für die Anwendung im realen Einsatzfeld geeignet.



(a) (b)

Abbildung 4. (a) Ergebnisse auf dem Datensatz 1. Beide Tracking-Systeme sind in der Lage, die Anzahl der in der Szene sichtbaren Personen mit einem leichten zeitlichen Versatz korrekt zu schätzen. Dies gilt sowohl für das Einfügen neu erscheinender Personen, als auch für das Löschen von verschwundenen Personen. Das selbst entwickelte Tracking-System, das mehrere Instanzen von CONDENSATION-Trackern verwendet, liefert immer ganzahlige Schätzungen, die der Anzahl der verwendeten Tracker zum jeweiligen Zeitpunkt entsprechen. Der Tracker von Tao liefert als Schätzung die mittlere Anzahl der Samples über alle Konfigurationen und somit einen gebrochenen Wert. (b) Ergebnisse auf dem Datensatz 2. Das System von Tao hat hier Schwierigkeiten beim Erscheinen der dritten und der vierten Person. Bei näherem Anschauen der Daten in Abbildung 3(b) erkennt man in den Bildern 55 und 72, dass das Gesicht besagter Person in den oberen Bildrand reicht und deshalb nur relativ wenig Hautfarbe sichtbar ist. Das selbst entwickelte System ist hier zuverlässiger, schätzt aber in den Frames 115 und 116 kurzzeitig eine Person zu viel. Dies ist darauf zurückzuführen, dass sich eine Person beim Verlassen der Szene sehr schnell zur Seite bewegt, so dass hier ein neuer Tracker initialisiert wird. Dieser wird aber sofort gelöscht, nachdem die Person verschwindet.

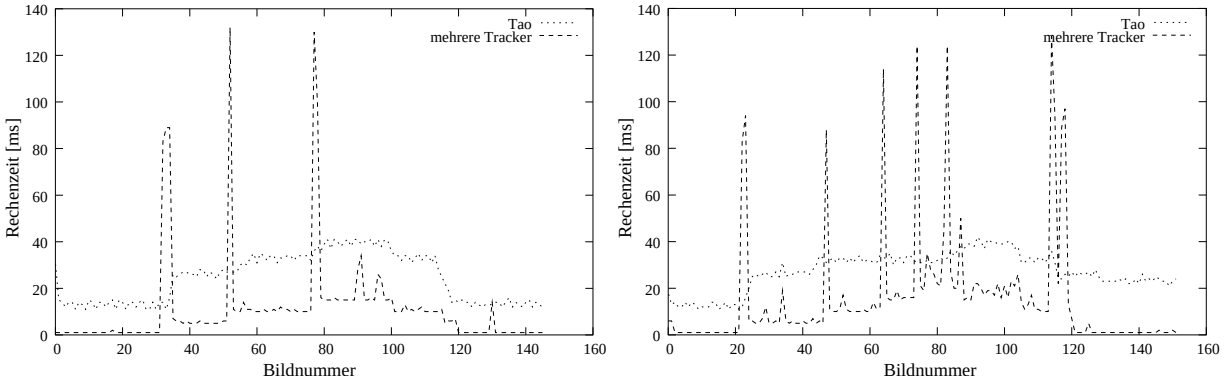


Abbildung 5. Rechenzeitbedarf der beider Verfahren auf den beiden Datensätzen. Die Rechenzeit pro Bild liegt beim Ansatz mit mehreren einzelnen Trackern unter der vom Ansatz von Tao. Die Ursache hierfür liegt darin, dass beim Ansatz von Tao das Eingangsbild in jedem Zeitschritt unterabgetastet wird, um die Anzahl der Hautfarbregionen im Bild zu schätzen, siehe Abschnitt 4.1.1, während beim selbst entwickelten Verfahren nur einzelne Pixel auf einem diskreten Raster betrachtet werden. Gegenüber dem relativ kontinuierlichem Verlauf der Rechenzeit bei Tao, entstehen beim selbst entwickelten Verfahren große Spitzen, wenn Instanzen neuer Tracker erzeugt werden.

6 Fazit

Bei den beiden Ansätzen handelt es sich auf den ersten Blick um sehr unterschiedliche Verfahren, die aber tatsächlich an vielen Stellen sehr ähnliche Berechnungsmodelle verwenden. Während bei Tao verschiedene Größen erst über die Wichtung von Sample-Konfigurationen berücksichtigt werden, gehen diese beim selbst entwickelten Verfahren direkt in die Konfiguration der Tracker ein. Daraus ergibt sich ein oft schnelleres Konvergieren der Tracker auf die tatsächliche Konstellation der Einzelbeobachtungen.

Literatur

[1] Fritsch, J., Lang, S., Kleinehagenbrock, M., Fink, G.A., and Sagerer, G. Improving adaptive skin color segmentation by incorporationg results from face detection. In *Proc. IEEE Int. Workshop on Robot and Human Interactive Communication (ROMAN), Berlin, Germany*, pages 337–343, 2002.

[2] Isard, M. and Blake, A. CONDENSATION – conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29(1):5–28, 1998.

[3] S. Feyrer and A. Zell. Detection, tracking, and pursuit of humans with an autonomous mobile robot. In *International Conference on Intelligent Robots and Systems (IROS ´99)*, pages 864–869, 1999.

[4] B. Schiele and A. Waibel. Gaze tracking based on face-color. In *Proc. Int. Workshop on Auto. Face and Gesture Recog., Zurich*, pages 344–349, 1995.

[5] Störring, M., Andersen, H.J., and Granum, E. Physics-based modelling of human skin colour under mixed illuminants. *Robotics and Autonomous Systems*, 34(3-4):131–142, 2001.

[6] Tao, H., Sawhney, H.S., and Kumar, R. A sampling algorithm for tracking multiple objects. In *Workshop on Vision Algorithms*, pages 53–68, 1999.

[7] Wilhelm, T., Böhme, H.-J., and Gross, H.-M. Automatischer Weissabgleich für eine omnidirektionale Kamera. In *Proc. 9. Workshop für Farbbildverarbeitung, Esslingen*, pages 43–50. Schriftenreihe ZBS, 2003.

[8] Wilhelm, T., Böhme, H.-J., and Gross, H.-M. Looking closer. In *Proceedings of the 1st European Conference on Mobile Robots*, pages 65–70. ZTUREK Research-Scientifi c Institute, 2003.

[9] Yang, J. and Waibel, A. Skin-color modeling and adaptation. *Lecture Notes in Computer Science*, 1352:687–694, 1998.