Multimodal Probabilistic Person Tracking and Identification in Smart Spaces

zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

der Fakultät für Informatik
der Universität Fridericiana zu Karlsruhe (TH)

genehmigte

Dissertation

von

Keni Bernardin

aus Karlsruhe

Tag der mündlichen Prüfung: **20.11.2009**

Erster Gutachter: **Prof. Dr. A. Waibel**

Zweiter Gutachter: **Prof. Dr. R. Stiefelhagen**

Lebenslauf

Keni Bernardin

Adresse:	Waldstr. 95 76133 Karlsruhe
Geburtsdatum:	14. Januar 1977 in Saarbrücken
Familienstand:	Ledig
Nationalität	Deutsch

Schule

1982 – 1988	Grundschule: Ecole St Jean l'Evangeliste, Port-au-Prince, Haiti
1988 – 1995	Gymnasium: College Canado-Haitien, Port-au-Prince, Haiti
Juli 1995	Abschluss: Abitur
November 1995	Anerkennungsprüfung für die Deutsche Hochschulzugangsberechtigung, Mönchengladbach

Universität

1996 – 2003	Studium der Informatik, Universität Karlsruhe
2001 – 2002	Vertiefungsfach KI und Robotik und Diplomarbeit an der Universität Tokyo, Japan
2002 – 2004	Wissenschaftlicher Mitarbeiter am Institute of Industrial Science, Universität Tokyo, Japan
Januar 2003	Abschluss: Diplom-Informatiker
2004 – 2009	Wissenschaftlicher Mitarbeiter am Institut für Theoretische Informatik, Lehrstuhl Prof. A. Waibel, Universität Karlsruhe.
seit Mai 2009	Wissenschaftlicher Mitarbeiter am Institut für Anthropomatik, Lehrstuhl Prof. R. Stiefelhagen, Universität Karlsruhe.

Karlsruhe, den 12.10.2009

Zusammenfassung

Intelligente Räume, die die sich in ihnen aufhaltenden Personen wahrnehmen und intelligente, Mensch-zentrierte Dienste anbieten, sind ein aktives Forschungsfeld. In diesem Kontext spielt das Tracken und Identifizieren von Personen (hier als "Identity Tracking" bezeichnet) eine wichtige Rolle, da es fundamentales kontextuelles Wissen liefert, welches für weitergehende Analysen des Verhaltens, der Aktivitäten oder der Interaktionen von Menschen verwendet werden kann. Dabei wird das Ziel verfolgt, mehrere Identitäten simultan zu verfolgen, während sie sich in der intelligenten Umgebung bewegen, unter Nutzung von unauffälligen Sensoren, wie z.B. weit entfernten Kameras und Mikrofonen. Ferner soll dies in Alltagsszenarien erreicht werden, in Umgebungen wie z.B. Besprechungsräume, Bürokomplexe, Wohnzimmer, usw., die wenige bzw. keine Einschränkungen des Verhaltens beobachteter Personen bedingen.

Eins der größten Hindernisse zur Realisierung des "Identity Tracking" in solchen realistischen Umgebungen ist, dass zuverlässige Merkmale zur eindeutigen Identifikation von Personen mit den oben genannten Sensoren nur schwer zu beobachten sind. Generell beobachtbare Merkmale, die mit dem groben Aussehen einer Person zusammenhängen, wie die Farbe der Kleidung, die Körpergröße, etc., sind oft nicht eindeutig oder ändern sich beträchtlich über die Zeit. Auf der anderen Seite sind invariante, stark personenspezifische Merkmale, wie z.B. solche, die durch Gesichts- oder Sprachidentifikation extrahiert werden, u.U. schwer zu erfassen (z.B. nur wenn eine gute frontale Sicht auf das Gesicht verfügbar ist, oder während die Person in einer Diskussion das Wort ergreift).

Der Lösungsansatz, der in dieser Arbeit verfolgt wird, besteht darin, aktiv und opportunistisch Personen zu identifizieren wenn eindeutige Merkmale erfassbar sind, und identifizierte Personen zu tracken bis weitere Beobachtungen möglich sind. Zwei Schwierigkeiten gilt es zu überwinden: Erstens sind einzelne Beobachtungen, die durch Gesichts- oder Sprachidentifikation gewonnen werden können oft unzuverlässig, da sie durch viele Störfaktoren beeinflusst werden. Es ist deswegen nötig, mehrere einer Person zuzuordnende Merkmale zu akkumulieren, um eine höhere Konfidenz bei der Identifikation zu erzielen. Zweitens kann nicht davon ausgegangen werden, dass das Detektieren und Tracken von Personen fehlerfrei realisiert werden kann. Personentracks können vertauscht werden, Tracks können verloren gehen, etc. Deswegen ist es notwendig, die Identitäten von Personen fortlaufend zu verifizieren und ggf. neu zu ermitteln.

Obwohl in der Literatur bereits einige Ansätze vorgestellt wurden, die sich mit dem Detektieren und Tracken von Personen in Sensornetzwerken, der Identifika-

tion von Gesichtern, der Sprachidentifikation, usw. befassen wurde noch kein umfassender Ansatz vorgestellt, der alle mit dem "Identity Tracking" in natürlichen Umgebungen verbundenen Probleme effizient angeht. Bestehende Ansätze befassen sich nur mit Teillösungen und sind nur in eingeschränkten Szenarien, für einzelne Benutzer oder für bestimmte Sensortypen oder -konfigurationen anwendbar.

In dieser Arbeit wird eine neue Methodik für das multimodale Tracken und Identifizieren von mehreren Personen unter Nutzung entfernt platzierter Mikrofone und Kameras vorgestellt. Im Gegensatz zu bestehenden Methoden integriert sie alle Teillösungen, die für das uneingeschränkte, audiovisuelle "ID Tracking" von mehreren Personen benötigt werden: Visuelles Tracking, akustische Quellenlokalisierung, Gesichtsidentifikation, Sprachidentifikation, Datenassoziation, Konfidenzbasierte Fusion, Erkennung unbekannter Personen, usw. Die Methode integriert sowohl starke, personenspezifische Merkmale zur Identifikation und Lokalisation, die für einzelne Personen nur spärlich über die Zeit erhältlich sind, als auch schwächere Merkmale, die zwar weniger eindeutig und akkurat, dafür aber regelmäßig, wenn nicht sogar kontinuierlich beobachtbar sind. Die Methode fusioniert weiterhin die akustische und die visuelle Modalität, sowohl für das Tracking als auch zur Identifikation, zur Erhöhung ihrer Flexibilität und Robustheit. Sie behandelt die Teilaufgaben der Lokalisierung und der Identifikation gleichwertig in einem probabilistischen Rahmenwerk, so dass Fehler in der Identifikation durch das Tracking ausgeglichen werden, und umgekehrt.

Der vorgestellte Ansatz, der "Joint Identity Tracking" (JIT) Filter, basiert auf dem Bayes'schen Filtern einer Vielzahl beobachteter Mermalstypen. Die Merkmalsextraktion wird mit "state-of-the-art" Algorithmen zur Detektion, Klassifikation und Identifikation realisiert. Die Fusion wird in einem Partikelfilteransatz realisiert, der speziell für die Behandlung von unregelmäßig auftretenden Beobachtungen, wie sie typischerweise in den angestrebten Szenarien auftreten, erweitert wurde. Weiterhin wird eine Methodik eingeführt, bei der die Identitäten aller bekannten Personen im Raum, und die Konfidenzen in deren Identifikation, gemeinsam probabilistisch ermittelt werden. Die Leistung des vorgestellten Verfahrens wurde systematisch evaluiert, auf einer umfangreichen audiovisuellen Datenbank, der "CLEAR Interactive Seminar Database", die in 5 verschiedenen intelligenten Räumen aufgenommen wurde. Diese Datenbank wurde schon in zwei internationalen Workshops, den CLEAR Workshops verwendet, um quantitative Benchmarks verschiedener Trackings- und Identifikationsteilaufgaben durchzuführen. Anzumerken ist, dass die schon bei den CLEAR Workshops verwendeten Metriken fürs Personentracking, sowie die hier neu vorgestellten Metriken fürs "Identity Tracking" im Rahmen dieser Arbeit entwickelt wurden.

Es wurde gezeigt, dass die Fusion verschiedener Modalitäten, selbst in schwierigen, unkontrollierten Umgebungen, und mit entfernt platzierten Sensoren, sich

vorteilhaft auf die Gesamterkennungsqualität auswirkt. Dies obwohl akustische und visuelle Beobachtungen, wie z.B. identifizierte Sprachsegmente und erkannte Gesichter, in der Regel asynchron auftreten und u.U. mit sehr unterschiedlicher Regularität für verschiedene Personen extrahierbar sind. Eine Object Tracking Accuracy ($MOTA$) von 77% und eine Identity Tracking Accuracy ($MITA$) von 81% konnten für den schwierigen CLEAR Seminar Datensatz erzielt werden.

Es wurden auch die Vorteile der zeitlichen Fusion von Identifikationsmerkmalen gezeigt, auch wenn die Assoziation von Beobachtungen zu automatisch generierten Personentracks unüberwacht und fehlerbehaftet geschieht. Durch kontinuierliches Tracking konnte die Identifikationsgenauigkeit erhöht werden. Gleichzeitig konnten durch die Wiedererkennung von Identitäten Fehler in der Initialisierung von Tracks vermieden werden.

Außerdem wurden die Vorteile einer probabilistischen, einheitlichen Integration aller Merkmale, lokalisierter sowie nicht-lokalisierter, akustischer sowie visueller, Trackings- und identifikationsmerkmale, auf globaler Ebene demonstriert. Dies ist in Hinsicht auf die erzielten Genauigkeiten, aber auch in Hinsicht auf das Systemverhalten, für den Fall dass einzelne Vorverarbeitungskomponenten, Sensoren, oder Modalitäten ausfallen. Es wurde gezeigt, dass die Leistung des JIT Filteransatzes graduell und relativ kontrolliert abfällt, so dass selbst bei extremen Ausfällen einzelner Merkmalsextraktionskomponenten, eine hohe Identity Tracking Accuracy beibehalten werden kann. So sinkt z.B. die MITA bei Ausfall aller visuellen Detektoren von 81% auf 67%, und bei Totalausfall aller Kameras auf 48%.

Der vorgestellte “ID Tracking” Ansatz wurde als echtzeitfähiges, verteiltes System in einem intelligenten Raum implementiert, bei dem zusätzlich zu fest fixierten Kameras und Mikrofonen, schwenkbare, automatisch gesteuerte Kameras verwendet wurden, um gezielte, aktive Gesichtserkennung zu betreiben. Dieses System stellt die erste Implementierung dar, bei der visuelles Tracking, akustische Quellenlokalisierung, Sprachsegmentierung und -identifikation, und aktive Gesichtssuche und -identifikation für mehrere simultane Nutzer in Echtzeit integriert wurden. Das System wurde mehrfach im Rahmen des Europäischen Projekts CHIL demonstriert.

Abstract

Smart spaces and environments that perceive their occupants' actions and offer intelligent human-centered services are an active topic of research. In this context, the tracking and identification of persons (referred to as "identity tracking") plays an important role, as it provides fundamental contextual knowledge upon which further analysis of activities or interactions can be performed. The overall goal is to simultaneously keep track of multiple identities evolving in the space using unobtrusive sensors, such as distantly placed cameras and microphones. Further, this is to be accomplished in everyday scenarios imposing little or no constraint on the natural behavior of users, such as in meeting rooms, office areas, living rooms, etc.

One of the main problems facing identity tracking is that in such realistic scenarios, reliable cues for person-specific identification are hard to obtain with the sensors described above. Generally observable features based on a person's overall appearance, such as the color of clothing, body height, etc., can be ambiguous (e.g. when all persons wear black) and may well vary considerably with time or environmental conditions (e.g. taking off one's jacket, sitting down). On the other hand, more invariant and person-specific features such as those gained by face or voice identification may only seldom be observable (such as when a good view of the face is available or when the person takes his or her turn speaking in a conversation).

The main idea followed in this thesis to overcome this problem is to actively and opportunistically capture reliable identification cues for each occupant whenever they become available and to keep track of identified persons until further observations can be made. This involves using focusable sensors such as steerable cameras to obtain high resolution facial close-ups and microphone arrays to determine the origin of speech. The difficulties to be dealt with are twofold: Firstly, single observations gained through face or voice identification are inherently noisy, being influenced by lighting conditions, low resolution, imperfect facial alignment, environmental noise, crosstalk, etc. This implies that identification cues need to be accumulated in time and multiple modalities should be used to increase the accuracy of identification. Secondly, in realistic scenarios, the tasks of automatically detecting and tracking persons in the first place cannot be assumed solved with perfect accuracy. Persons may be missed, tracks may be confused or lost. This means that person identities need to be correctly recovered when observations again become available. While some amount of work has been done on the field of tracking and identification using sensor networks with overlapping or even non-overlapping views, none of the approaches so far tackle all the related problems efficiently. Most integrated approaches

rely on general appearance features, such as color, and build on the assumption that features for identification are jointly available with features for tracking with every observation made. While some approaches use person-specific features such as provided by face identification, they still rely on the continuous availability of high resolution face images in very restrictive setups. Approaches that use acoustic features for identification typically assume that the number of persons is known a priori, that speakers take frequent turns, and do not keep track of their locations except in very restrictive setups. More importantly: Almost all approaches found in the literature that target multiple users are limited to applications where the detection (and spatio-temporally local tracking) of persons can be realized flawlessly and build on the results of this step for identification. A better approach would be to integrate identification and tracking into one framework, such that errors in tracking are less detrimental to overall accuracy.

In this thesis, a new methodology is introduced for the multimodal tracking and identification of multiple persons by seeking and integrating reliable ID cues whenever they become observable. The method opportunistically integrates person-specific identification cues that can only sparsely be observed for each person over time and keeps track of the location of identified persons while ID cues are not available. It also fuses the acoustic and visual modalities to increase its robustness and flexibility and probabilistically integrates tracking and identification at the same level, such that errors made by one are compensated by the other. Finally, it represents a general framework for ID-Tracking in that it supports different types and configurations of sensors, different modalities with varying reliability or availability, and is scalable to different rooms, sensor setups, etc. The developed method is a generative approach based on Bayesian filtering of high level tracking and ID cues. It is implemented as a particle filter which approximates the probability density of the persons' presence, locations and identities by a set of samples or particles. The belief of the particle filter is propagated in time and updated which each new observation concerning person locations or identities. In this thesis, a new formulation is developed to represent the belief of the particle filter about the state of the world (the person locations and identities), and to keep the state space tractable while avoiding to rely on flawless detection and tracking results. Moreover, the formulation allows the integration of sparsely available observations concerning identities together with tracking-related observations, which arrive at a constant and high framerate.

The performance of the developed system is evaluated in the context of small meetings with several users taking place in "smart" rooms, and the effect of using single or multiple modalities on identification accuracy is investigated. Experiments have been made using a large multimodal database of recordings captured in 5 different instrumented rooms. This database, which was used in two international evaluations, the CLEAR evaluation workshops, allowed a

thorough, quantitative evaluation of the approach. Note that the accuracy metrics used in the CLEAR workshops themselves were developed in the course of this thesis in an effort to provide a clear and generally applicable methodology for the quantitative evaluation of multiple target tracking performance. Experiments involving real-time identity tracking were also made with a distributed implementation of the approach using several cameras and microphones in a smart room. The results show that the integrated approach is robust to tracking failures and degrades gracefully with decreasing tracking accuracy. They also show that the fusion of audio and visual modalities can help achieve noticeable identity tracking accuracies, even in relatively uncontrolled situations with multiple persons, occlusions, cross-talk, etc., and using available state-of the art tracking, face and voice identification components.

Contents

1 Introduction **11**
- 1.1 Motivation . . . 11
- 1.2 Challenges and Thesis Contributions . . . 11
- 1.3 Thesis Overview . . . 16

2 Related Work **17**
- 2.1 Multi-Sensor Audio-Visual Person Tracking . . . 17
- 2.2 Face Identification, Speaker Identification and Multimodal Fusion 20
- 2.3 Identity Tracking . . . 26
- 2.4 Performance Evaluation . . . 29

3 Features for Identity Tracking **33**
- 3.1 Overview of Feature Types . . . 35
- 3.2 Features for Tracking . . . 38
  - 3.2.1 Foreground features . . . 38
  - 3.2.2 Detection features . . . 39
  - 3.2.3 Color features . . . 41
  - 3.2.4 Top View Features . . . 43
- 3.3 Features for Identification . . . 46
  - 3.3.1 Classical Identification Tasks and Metrics . . . 46
  - 3.3.2 Features Extracted from Face Identification . . . 49
  - 3.3.3 Features Extracted from Voice Identification . . . 53
- 3.4 Spatial Localization and Composition of High-Level Features . . 54
  - 3.4.1 Visual Localization . . . 55
  - 3.4.2 Acoustic Localization . . . 58
  - 3.4.3 High Level Feature Description . . . 59
- 3.5 Summary . . . 60

4 Probabilistic Multiple Identity Tracking using Irregularly Observable Features **61**
- 4.1 Localization . . . 63
  - 4.1.1 Particle Filter Framework . . . 64
  - 4.1.2 Observation Models for Scoring . . . 66
  - 4.1.3 Prediction and Resampling . . . 72
  - 4.1.4 Mutual Track Exclusion . . . 72
  - 4.1.5 Occlusion Handling . . . 74

4.1.6 The Issue of Observability . . . 78
4.1.7 Correcting the Prior on Uncovered Observations . . . 81
4.1.8 Track Creation and Deletion . . . 83
4.1.9 Unsupervised Color Model Learning . . . 84
4.1.10 Output Track Locations . . . 86
4.2 Identification . . . 87
4.2.1 Warping Confidence Values . . . 90
4.3 Joint Identity Filtering . . . 91
4.3.1 Assumptions and Task Definition . . . 91
4.3.2 Joint Update in Identity Space . . . 93
4.3.3 Person Creation, Deletion and Data Association . . . 95
4.3.4 Evaluating Identities . . . 96

5 Performance Metrics **99**
5.1 Performance Metrics for Multiple Object Tracking . . . 100
5.1.1 Establishing Correspondences Between Objects and Tracker Hypotheses . . . 101
5.1.2 MOT Metrics . . . 105
5.2 Performance Metrics for Identity Tracking . . . 108
5.2.1 Multiple Identity Tracking Accuracy . . . 110
5.3 Evaluating Open Set Identification Performance . . . 111

6 Experimental Evaluation **115**
6.1 Evaluation Database . . . 115
6.2 Experimental setup . . . 118
6.2.1 Visual Recognition . . . 119
6.2.2 Acoustic Recognition . . . 122
6.3 Baseline . . . 124
6.4 Evaluation Results . . . 126
6.4.1 Thresholding Identification Confidence . . . 128
6.4.2 Identity Tracking Accuracy . . . 129
6.4.3 Baseline Comparison . . . 132
6.4.4 Modality Fusion . . . 136
6.4.5 Temporal Fusion . . . 139
6.4.6 Graceful Degradation . . . 142
6.5 Discussion . . . 144
6.6 Live System . . . 144
6.6.1 Active Camera Face Capture . . . 146
6.6.2 Speech Detection and Recognition . . . 148
6.6.3 Target and Camera Selection Strategies, Recognition of Standard Events . . . 148
6.6.4 Experiments using the Live System . . . 149

7 Conclusion **153**

7.1 Summary and Discussion 153
7.2 Future Directions 155

Own Publications **161**

References **163**

List of Figures

3.1 Overview of features for identity tracking. 37
3.2 Example foreground support map in a camera view of a CLEAR Seminar recording. 39
3.3 An upper torso detection in a corner camera view. 40
3.4 An example of learned upper body color models for multiple subjects. 42
3.5 Examples of features extracted from a top view overlooking the smart space. 45
3.6 An example ROC curve for an open set identification case. . . . 50
3.7 Local appearance-based face recognition 51
3.8 The estimated torso area relative to a detected face. 56
3.9 The uncertainty in the 3D location of a detected person. 57

4.1 The inner and outer bounding boxes used to estimate the foreground likelihood score for a particle. 69
4.2 The occupancy maps calculated upon the grid-based discretization of the smart space. 74
4.3 The computation of occlusion factors based on grid discretizations. 76
4.4 The extension of the grid-based occlusion model to include height information. 77
4.5 The grid-based occlusion maps for three different corner cameras, as seen from a same top view point. 78
4.6 The output of the localization step. Again, track-specific occupancy grids are used to efficiently cluster particles and infer a locally smoothed average. 87
4.7 The output of the integrated identity tracking system. 97

5.1 Mapping tracker hypotheses to objects. 102
5.2 Optimal correspondences and error measures. 104
5.3 Computing error ratios. 107
5.4 The different types of errors that can be produced in identity tracking. 108
5.5 Example curves for localization accuracy (LA), identification accuracy (IA) and Multiple Identity Tracking Accuracy ($MITA$). 111
5.6 An example ROC plot for the CCR, FCR (and implicitly the FRR) in relation to the FAR in the case of identity tracking. . 113

6.1 Scenes from the CLEAR 2007 Interactive Seminar database. . . 116
6.2 The ROC curves for the single frame open set face recognition performance ($k = 10$ to 100). 121
6.3 The ROC curves for the single frame open set face recognition performance ($k = 5$ to 10). 121
6.4 The ROC curves for open set speaker recognition on individual one second sequences. 124
6.5 The process of mapping localized ID cues to person tracks in the baseline system. 125
6.6 Multiple identity tracking performance of the JIT approach on the CLEAR'07 dataset. 127
6.7 The $MITA$ obtained with thresholding of identification cues, based on their normalized, warped confidence scores. 128
6.8 The evolution of the $MITA$ as a function of time. 130
6.9 The evolution of component ratios as a function of time. The FCR and FPR are close to zero and are not plotted. 131
6.10 The ROC curves for the joint identity tracking task on the CLEAR'07 seminar data. 131
6.11 Identification accuracies for the JIT and the baseline approach in the presence of component failure. The identification accuracy for the baseline system in the case no tracking is possible (*Base_NoTracks*) is always zero, which is why the corresponding curve is not visible here. 133
6.12 Localization accuracies for the JIT and the baseline approach in the presence of component failure. The localization accuracy for both systems in the case no tracking is possible ($NoTracks$) is always exactly zero. 134
6.13 The $MITA$ scores for the JIT approach and the baseline system. 135
6.14 $MITA$ scores for single modality acoustic or visual and for multimodal identification. 137
6.15 ROC curves for single modality and for multimodal identity tracking. 138
6.16 Evolution of the frame-based Multiple Identity Tracking Accuracy (MITA) in time. 139
6.17 A comparison of MIT scores reached in the multimodal case, when using the sum rule, the product rule, or without separation of identity models. 140
6.18 The ROC plots for comparison of the single model strategy, the sum rule and the product rule for multimodal identity tracking. No significant difference can be observed. 140
6.19 The effects of confidence based temporal fusion on identity tracking accuracies. Both the MIT scores and the open set ROC performance are significantly increased when temporal fusion is made. 141

6.20 Overview of the components of the live multimodal identity tracking system. 145
6.21 Face detection and alignment in active camera images. 147
6.22 Examples of standard events detected using the knowledge about person locations. 150
6.23 An example evaluation scenario for the live identity tracking system. 151

List of Tables

6.1 Person tracking performance for the JIT approach on the CLEAR'07 dataset. 126
6.2 Highest reached site-specific and global identity tracking scores for an acceptance threshold $Th_{known} = 0.3$. 129
6.3 Person tracking performance for the purely acoustic, the purely visual, and for the audiovisual cases. 136
6.4 Person tracking and Identity Tracking performance in the presence of individual component failure for the multimodal case. . . 142
6.5 Person tracking and Identity Tracking performance with varying numbers of particles per track. 143

1 Introduction

1.1 Motivation

In the context of smart environments, the ability to track and identify persons is a key factor, determining the scope and flexibility of analytical components or intelligent services that can be provided. While some amount of work has been done concerning the camera-based tracking of multiple users in a variety of scenarios, technologies for acoustic and visual identification, such as face or voice ID, are unfortunately still subjected to severe limitations when distantly placed sensors have to be used. Because of this, reliable cues for identification can be hard to obtain without user cooperation, especially when multiple users are involved.

In this thesis, a novel technique is presented for the tracking and identification of multiple persons in a smart environment using distantly placed audio-visual sensors. The technique builds on the opportunistic integration of tracking as well as face and voice identification cues, gained from several cameras and microphones, whenever these cues can be captured with a sufficient degree of confidence. A probabilistic model is used to keep track of identified persons and update the belief in their identities whenever new observations can be made.

Possible application areas for identity tracking are:

- Meeting summarization
- Lecture browsing
- Smart reactive/proactive environments
- Surveillance and security

1.2 Challenges and Thesis Contributions

One of the main problems on the way to achieving unobtrusive multiple user identity tracking is that in realistic scenarios, reliable cues for person identification are hard to obtain with conventional, distantly placed cameras and

microphones. Even when the detection and tracking of persons can be managed to a satisfactory extent, finding features that allow to identify them or distinguish them from one another can be problematic.

First of all, generally observable features based on overall visual appearance, such as the color of clothing, the direction and speed of movement, body height, etc., are often ambiguous. In a business meeting in which most attendees wear black suits, for example, the color of clothing will not be of much use for telling them apart. As for the person height: When tracking with distant cameras, the uncertainty associated with the estimation of the person height can well exceed the often small difference in actual height, even assuming persons are standing perfectly upright. General appearance features may also vary considerably with time or environmental conditions. The perceived color of clothing can change dramatically as a person walks through a scene, due to illumination conditions or, in a much more pragmatic case e.g., when he or she takes off his or her jacket. It can also be quite different from camera to camera, as differences in the viewed scene or in the camera sensor itself don't always allow the assumption of color constancy. The person height or shape changes dramatically with posture, for example when a person sits down or stands up. Likewise, features based on speed and direction of motion build on the assumption of some form of regularity underlying the dynamics of tracked objects. This is, however, rarely true for humans. In fact, the converse often holds. A common scenario would show a person standing or sitting still for long periods of time, suddenly starting to move around, stop, for example at a door to let someone else in, turn around, change direction to avoid an obstacle, and so forth.

Compared to general appearance features, with limited constancy and discriminative power, face or voice identification features can be quite invariant and person-specific. The problem here is that they may only seldom be observable. A good view of a person's face, e.g, may be hard to obtain in non-restrictive scenarios where no fixed person location or general direction of attention can be assumed. Although much progress has been made on handling the effects of shadows, uneven illumination or even partial occlusions of the face, today's state of the art face identification techniques still specialize on frontal views with relatively high resolutions [121]. It is easy to see why the application of such techniques is not straightforward in the scenario, for example, of a small group meeting observed using a camera mounted on one of the meeting room walls. Many of the attendees' faces are likely turned away from view, sometimes for lengthy periods of time, as they are looking at each other, or even downward at their notes, making no effort to explicitly look into the observing camera. Even when a face is visible, it will often be of very low resolution, making an identification difficult. Of course, these problems can be overcome e.g. by multiplying the amount of observing cameras, placing cameras directly on the meeting table in front of every participant, or even using a panoramic camera placed in the middle of the table [22]. All these solutions, however, which aim at improving

the coverage of interesting regions of the observed smart space, come at the cost of more specialized and expensive installations or more restrictions on the freedom of the users. If face identification is to be performed in common meeting rooms, lecture rooms, offices or living rooms, without large amounts of installed sensors, predefined seating positions, etc., one has to expect that usable frontal views of some of the observed persons' faces can often not be captured.

The limitations of speaker identification in this context come from the nature of human communication itself. When a small group of persons is engaged in conversation, they usually take turns speaking, which automatically limits the availability of acoustic features for any one participant. In extreme cases, a participant in a small meeting may never say a word at all during the whole session. Interestingly, the inverse problem may be even more severe: Crosstalk, which occurs when several persons are speaking at the same time, still represents a tough problem for state of the art speaker identification techniques. It requires the use of special techniques, e.g. blind source separation , to separate the overlapped speaker signals before analysis can be performed. Unfortunately, crosstalk can be very common in natural scenarios with many users, as several subgroups often form, each with their individual conversation threads. The problem is made even more acute through the addition of other noise sources, such as printers, beamer fans, clapping, laughter, and so forth, which all affect the speech signal to be analyzed. The result is that a reliable identification for a given user, based on his speech, may only be possible for a few rare moments in time, if no dedicated hardware, such as lapel or head set microphones, is to be used.

Apart from face or voice identification techniques, delivering strong person-specific information, there is also the possibility of using specialized markers or devices, such as RFID tags, which are to be worn by the users of the smart environment. There is quite an amount of research in the domain of ubiquitous computing, which deals with the localization and identification of multiple users in wide areas using such wearable devices [39]. The advantage is that, compared to other techniques, they deliver relatively constant, reliable and user-specific information. The obvious downside is that users are required to wear these devices in the first place, which may be perceived as intrusive, and in some cases is just not practical or even feasible. An example would be a small seminar organized on-the-fly, involving a presenter and few students. Here, the overhead of keeping wearable devices ready for each possible participant and the intrusive requirement to put them on at the start of the session can make this option much less attractive than observing the scene with a few pre-installed and discretely placed cameras and microphones.

The goal that is pursued here is to achieve unobtrusive identity tracking using distantly placed sensors, overcoming the difficulties posed by face and voice identification in natural, multiple-user scenarios. The main idea followed is to

opportunistically integrate reliable identification cues for each person whenever they become available and to keep track of identified persons until further observations can be made.

The difficulties to be dealt with are twofold: Firstly, single observations gained through face or voice identification are inherently noisy, being influenced by lighting conditions, low resolution, imperfect facial detection and alignment, environmental noise, crosstalk, etc. Therefore, to increase the accuracy of identification, single observations need to be accumulated for each individual and their identity estimated using the sequence of observations made so far. In multiple user scenarios, in which individuals move around freely and change their location frequently, this can only be accomplished by keeping track of their positions, such that each new observation can be correctly associated to the right person. Another consequence is that multiple independent modalities should be used whenever possible to alleviate the effects of missing or noisy observations. If a person's face is not visible for longer periods of time, for example because he or she is turned away from the camera, we have to rely on his or her voice alone for identification. In this sense, the two modalities are not simply fused to increase accuracies when both are available. Rather, they act in a truly complementary fashion.

The second difficulty is that in realistic scenarios, the tasks of automatically detecting and tracking persons in the first place cannot be assumed solved with perfect accuracy. Even using state of the art systems, in natural indoor scenarios with a sufficient amount of clutter, persons may be not be accurately detected, false tracks may be wrongfully initialized, tracks may get confused or lost, etc. This means that person identities need to be correctly recovered when observations again become available. This also means that considering tracking information itself as inherently noisy can help avoid certain types of errors. An example scenario is one where three different voices have been reliably identified in the smart space, although only one person track could be initialized. In this case, using the information gained from voice identification to estimate the number of persons present leads to the correct result. In many cases, however, the converse will be true, such that all sources of information, person detection and tracking, visual and acoustic identification need to be carefully balanced.

While some amount of work has been done on the fields of audio-visual tracking and identification using sensor networks, none of the approaches so far tackle all the related problems efficiently. Most integrated approaches rely on general appearance features, such as color (or on RFID tags and other worn sensors), and build on the assumption that features for identification are jointly available with features for tracking with every observation made. Moreover, approaches that attempt biometric identification of multiple users in smart spaces, using faces or voices, often neglect the difficult problems of data association and temporal fusion of identification cues. More importantly: Almost all approaches found in

the literature that target multiple user tracking and identification are limited to applications where the detection and tracking of persons can be realized flawlessly and build on the results of this step for identification.

In this thesis, a novel method is presented for the multimodal tracking and identification of multiple persons by fusing reliable tracking and ID cues whenever they become available. The method:

- Opportunistically integrates person-specific identification cues that can only sparsely be observed for each person over time
- Keeps track of the locations of multiple identified persons while ID cues are not available
- Combines the acoustic and visual modalities to increase its robustness and flexibility
- Does not rely on accurate detection and tracking, but rather considers both a person's location and identity as attributes to be estimated.

The developed method is a non-parametric approach based on sequential Bayesian filtering of various types of tracking and ID cues. It estimates the probability densities of a person's presence, location and identity based on the sequence of observations made. The proposed approach has been tested on a large annotated audio-visual corpus, the CLEAR Seminar Database [104; 103], comprising a total of 200 minutes from 20 different recorded small meetings. This database, captured in smart rooms using distantly placed sensors, features visual streams from several cameras on which tracking and face identification can be performed, as well as audio streams from several microphone arrays for speaker tracking and identification.

The main contributions of this work are the following:

- It presents a novel framework for the simultaneous tracking and identification of multiple users in realistic, unconstrained and uncooperative scenarios using only distantly placed sensors. This means that no dedicated cameras on meeting tables, no close talking microphones, no restrictive setups and no person markers are used. The associated problems of observability and data association make this a very difficult task which, to date, has not been tackled to this extent in the literature.
- It presents a method for the opportunistic fusion of observations with variable availability, specificity and accuracy. General appearance features such as color and upper body shape, as well as more specific features such as those gained through face and voice identification are integrated in a general, theoretically sound framework. Most related work on multimodal fusion assumes observations come in a synchronous way, and are continuously available.

- It represents the first integrated approach to combine acoustic localization, visual tracking, face identification and distant speaker identification for multiple persons in an open set identification scenario. While a lot of research has been done on the individual components, only partial work was made on integration, and no fully integrated system, tackling all the associated problems has been proposed.
- It is a robust, scalable and real-time capable method, which degrades gracefully with individual modality failure. It is generalizable to different types of sensors, sensor configurations, smart space sizes, and so forth.

Additional contribution are as follows:

- The presented method combines intelligent fusion with active seeking techniques through the use of automatically steered active cameras
- It is the first to effectively combine acoustic speaker tracking and identification for multiple persons using only far-field microphones.
- It presents a novel set of metrics for the evaluation of multiple object tracking performance. These Multiple Object Tracking (MOT) metrics provide a clear and generally applicable methodology for quantitative evaluations on large datasets and were already used in several international evaluations, as well as in a growing number of independent publications.

1.3 Thesis Overview

In the following chapter, an overview of the related work on audio-visual person tracking, audio-visual identification, identity tracking and performance evaluation is given. Chapter 3 then describes the algorithms and classifiers used to extract the low-level and high-level features used as inputs for the probabilistic identity tracking. Chapter 4 presents the proposed joint identity tracking algorithm, the JIT filter, and Chapter 5 explains the newly proposed metrics used for its evaluation. In Chapter 6, the proposed identity tracking approach is thoroughly evaluated, as well as compared to a baseline system which performs fusion in a sequential way, by building on the results of a tracking step to infer identities. A real-time capable implementation of the JIT filter approach is also briefly presented. Finally, Chapter 7 gives a summary and an outlook to future directions of research.

2 Related Work

This section presents the state-of-the-art in audio-visual person tracking and identification for smart environments. First, an overview of multi-sensor multi-person tracking techniques is given. Then, current approaches for visual face identification, acoustic speaker identification, as well as techniques for their multimodal fusion are presented. An overview is given on the field of "identity tracking" for smart environments, i.e. of techniques that combine visual or acoustic tracking and identification components in order to unobtrusively identify and keep track of variable amounts of persons. Finally, as one of the main contributions of this work is the design and evaluation of novel metrics for identity tracking performance, a short review of the field of tracking performance evaluation is also given.

2.1 Multi-Sensor Audio-Visual Person Tracking

The detection and tracking of persons in indoor environments has attracted an increasing amount of attention in the fields of computer vision and signal processing. One of the most influential early works was the PFinder by Wren et al. [114], but many more approaches were developed in the past decade, using a wide range of sensors end techniques, for indoor and relatively constrained outdoor environments [79; 45; 33; 34; 32; 71; 30; 51; 61; 120]. While earlier techniques were designed to track at most one person in a static, controlled environment, theoretical and algorithmic advances, as well as the constant increase in processing performance, have lead to the emergence of techniques for the automatic detection and simultaneous tracking of a variable amount of persons in relatively cluttered scenes, through occlusion, etc.

While purely vision-based techniques have been explored for quite some time, recently approaches that combine acoustic source localization and visual tracking have been proposed [123; 111; 25; 27; 85; 21; 40; 57]. Apart from increasing the robustness of visual trackers in the cases the target person is speaking, audio-visual localization techniques offer the advantage of exploiting the complementary nature of both modalities, such that tracking can be pursued, e.g., using only sound while a person is visually occluded.

One of the earliest approaches was proposed by Zotkin et al. [123]. The authors use a pair of steerable PTZ cameras and a pair of microphone arrays to audio-visually track alternating speakers in a frontal setup. They present a probabilistic framework for the integration of modalities, based on particle filters. It integrates skin color features, face detections, and Time Delays of Arrival (TDoAs) from the microphone pairs. It implements a simple occlusion handling mechanism to deal with partial measurements due to occlusion of the tracked object from one of the cameras, or to missing TDOA estimations due to noisy or weak audio channels. Though the scene may be populated by multiple users, the approach performs tracking for one person at a time and automatically switches to the last active speaker. The authors show that by integrating both modalities into one probabilistic framework, they can increase the accuracy of person tracks, bridge gaps in individual modality observations, and detect alternating speakers, realizing a speaker diarization based on known person locations.

Similarly, Vermaak et al. [111] also propose the use of particle filters for audio-visual tracking of speakers in a video telephony scenario. They use one static camera to extract the silhouette of users' heads and shoulders and a pair of omnidirectional microphones for speaker localization. Although their setup is rather constrained, it allows for multiple user interaction, as tracking automatically switches to the current active speaker. The state of the tracked target is modeled in each particle by a configuration comprising the image location and an affine transformation for the head-outline template. The authors note the advantages of of audio-visual tracking, as tracks that are lost due to fast movement or occlusion in the visual case are recovered by the audio modality. The system is still limited to tracking one user at a time, though. Generally, while single camera setups can be useful for videoconferencing applications, etc., multiperson settings, as occur in seminars or meetings require the use of multiple cameras and microphones to cover an entire observation space (the table, entrance area, whiteboards, etc.)

Checka et al. [25] present a system that combines multiple audio and visual sources to track multiple persons in a cluttered scene. In contrast to the previous approaches, their system uses distantly placed cameras and microphones observing a large space, a setup which is much closer to the concept of smart perceptive spaces followed here. Again, a particle filter approach is chosen, with the state space including the number of persons present, their location, and whether each person is talking. They use a visual whole body appearance model based on foreground segmentation and an acoustic model based on the short time Fourier transforms of the microphone audio signals. The audio-visual observation likelihood is obtained by multiplying individual likelihoods. They evaluate their system on short audio-visual sequences involving two to three persons walking on predefined paths with little occlusion, with strict turn-taking dialogs, and with manual initialization. They obtain good results for track-

ing and speaker determination and again show the robustness and flexibility of particle filters in the presence of noise.

In [57], Khalidov et al. present a system for the audio-visual tracking of speakers in meetings, using a human-like perceptual setup. A stereo camera system and a binaural microphone array are placed on the table, observing a group of 2-3 participants. The 3D locations and speech activity of participants are found by clustering of the audio-visual observations, and by an Expectation-Maximization-like inference mechanism. Here, the system builds on the complementary nature of the audio and visual modalities to gather sufficient clustering evidence and utilizes the Bayesian Information Criterion (BIC) to determine the number of persons present. The limitation of such systems, of course, is the frontal nature of the sensor setup, which limits the amount and freedom of users that can be observed.

Gatica-Perez et al. present a much more elaborate setup in [40]. Their observation space is composed of 3 uncalibrated cameras offering frontal views of participants seated around a meeting table and of the presentation area. It also comprises an 8-element circular microphone array. Again, visual and acoustic observation models are defined using head ellipsoid model fitting, skin color blobs and the Generalized Cross Correlation (GCC) between signals from microphone pairs. The observation likelihood is defined as the product of individual model likelihoods. Filtering of speaker positions is done using the Markov Chain Monte Carlo (MCMC) technique. As in the particle filtering technique, the target state (person locations and speaker activity) is approximated by a set of samples. In contrast to standard particle filtering, though, the MCMC sampling technique allows to jointly track several objects in a tractable manner, while preserving the rigorous joint state-space formulation. The determination of the number of person present in a camera image is not, as e.g. in [25], made in the MCMC framework, but in a separate process based on skin color blobs and creation and deletion regions where new tracks may be initialized. Further, the approach uses non-overlapping frontal views and tracks only the smart space occupants which are in the field of view of a camera.

Another dimension in tracking is reached in approaches by Nickel et al. [85], Brunelli et al. [21], Lanz et al. [63], Bernardin et al. [4] and Ferrer et al. [24]. In these approaches, speakers in a smart seminar room are tracked using a variety of audio-visual features such as foreground support, color, shape and appearance, edges, and audio signal correlations from a number of distantly placed cameras and microphone arrays. They build on the particle filter formulation to flexibly integrate various types of observations and effectively track in the presence of noise and clutter. While [85] and [21] are designed for the tracking of single persons (the seminar presenters) through noise and clutter in the form of other moving persons in the audience, changing lighting conditions, reverberations, etc., the techniques presented in [63; 19; 24] are able to simultaneously

track several users, initialize new tracks automatically and adapt person models on-the-fly. The difficulty in multi-person tracking under these conditions is that heavy occlusions in the camera views as well as noise and cross-talk are frequent. Persons need to be tracked as they move freely around the smart space, sit at the table, take turns speaking, laugh, pass near each other, etc. A special probabilistic formulation for handling of occlusions under these conditions and for jointly managing several persons while keeping the state space tractable is presented in [62; 63]. There, it is shown that specifically modeling the interactions and occlusions between tracks can greatly improve performance. Although [63; 24] are visual systems, which integrate the audio features in a post-processing step to infer the active speaker, their extension to include acoustic features in the probabilistic estimation is straightforward. These systems were evaluated in the international CLEAR evaluations [106; 105] on large datasets of realistic recordings featuring small interactive meetings. The results presented show that in the case of single person tracking, in those scenarios, very high accuracies can be achieved. The fully unconstrained multi-person scenario, in contrast, still poses great challenges to tracking techniques, though respectable accuracies can already be reached. These results, as well as a further analysis of the accuracies obtained in the CLEAR person tracking tasks, show that for such realistic and challenging scenarios, the flawless detection and tracking of multiple occupants is still not a realistic prerequisite.

The above presented approaches, just as many others, exemplify why the particle filter framework has become the method of choice for tracking multiple targets under extreme conditions. Since the introduction of the Condensation algorithm by Isard and Blake [52], particle filters have gained a steady increase in popularity and many variations have been proposed, for application in various fields [71; 72; 108; 78; 112; 62; 75; 86], to cite just very few examples. Compared to other sequential Bayesian estimation techniques, the advantage of particle filters lies in their flexibility with respect to the types and numbers of features they support, their robustness in the presence of noise, and in the non-parametric fashion in which they represent the belief about the target state, which makes them applicable for highly non-linear, non-Gaussian estimation problems.

2.2 Face Identification, Speaker Identification and Multimodal Fusion

Zhao et al., in [121] give a comprehensive review of research in the domain of face recognition. They summarize the face identification task as made up of three main steps: 1) Face detection and rough normalization, 2) Facial feature extraction and accurate normalization, 3) Identification and/or verification. This

division also reflects the main problem associated with face identification in natural, smart environments, when explicit user cooperation can not always be expected and sensor placement can not always guarantee a good view of faces: The first two steps in classical face identification become a major challenge for which no satisfactory solution has yet been proposed. Indeed, while applications are quite varied, the vast majority of face recognition research and the major evaluation efforts in this domain have focused on scenarios where faces are recorded at relatively close distance. Examples are the Face Recognition Grand Challenge [89], the FERET database [89] or the AR database [73]. For these conditions, many approaches have been presented, to tackle the still difficult problems of uneven illumination [10; 42; 110; 26], pose changes [42; 29], occlusion [74; 38; 115], etc.

In the case of surveillance or smart space applications, the size of faces that can be observed plays a major role. Face resolutions are typically much lower (down to 15×15 pixels, or less) than in close-up identification tasks, such as biometric verification, HCI, or videoconferencing (typical sizes are, e.g., 64×64 or 128×128 pixels). At low resolutions, facial features can no longer be accurately determined, such that the alignment step for face identification becomes difficult. Even the detection itself can pose a problem, especially when uncooperative users, shadows, low contrasts, partial occlusions by the hands, etc, render the task more difficult. Standard face detection algorithms, e.g. the approach proposed by Viola and Jones [113] can typically detect frontal faces at sizes down to of $24x24$ pixels. Beyond this, special techniques have to be used, involving a combination of features and steps [83]. Finally, the identification step itself becomes more difficult, as fewer features can be observed to help discriminate between persons.

As already proposed in [121], a method to deal with these problems is the use of video. Using a sequence of observations from, e.g., tracked faces helps alleviate the detection problem. It also poses an advantage for recognition, as shown e.g. by Stallkamp et al. in [98]. Here, recognition is performed in an unaware fashion for individuals entering through the door of a smart space. Features are extracted through a combination of face detection, skin color-based tracking, and eye detection and tracking, such that the identification decision is made at sequence level using all available observations. Frame level identification scores are summed up using both a simple sum rule and various weighted summation techniques, based on frame level ID confidences. The authors show that the temporal fusion of scores in video-based identification brings a dramatic improvement, compared to frame-based identification.

Methods that build on video-based fusion at extremely low resolutions have also recently been proposed [101; 20; 68; 36]. In [101], Stergiou at al. present a system that combines PCA-based nearest neighbor classification, subclass LDA-based classification and Bayesian face recognition [80]. The observations for

classification come from four distant video streams of a smart meeting space, where faces of target persons have been manually labeled beforehand. No facial feature detection or alignment is performed. Faces are resized to 32×48 pixels and fed to the different classifiers. Identification is made at the sequence level by fusing first in the temporal domain and across cameras for each classifier type individually, and then at the classifier level, using confidence scores derived from frame-level identifications. Similarly, Ekenel et al. in [36] present a system for the temporal fusion of frame level face identifications across cameras. In contrast to [101], they utilize a classification technique based on the local modeling of appearance features. Local appearance modeling has shown to be relatively robust to illumination changes, occlusion, etc, which makes it well suited for applications in realistic environments. The authors downscale cropped faces to a size of 20×20 pixels, subdivide the resulting patch into a regular grid of blocks, and locally extract normalized DCT (Discrete Cosine Transform) coefficients for each block. The classification is made using a nearest neighbor classifier, with identification confidences, based on the difference between the two nearest neighbors, used in temporal fusion. In all these cases, the confidence-based temporal fusion of identification scores has brought great improvements in performance. It was also shown that recognition accuracies increase with the length of the observed sequence. Overall, recognition accuracies of over 90% could be reached in a closed set task with 28 subjects, even at extremely low resolutions. One must remember, though, that the task of *detecting* the faces automatically was not tackled in any of these approaches, and all accuracies are computed for the case of manually annotated face bounding boxes only.

The automatic identification of speakers based on their voice can be a key feature of smart environments, and a complementary modality to face identification. The advantage of the audio modality is the larger coverage offered, in contrast to video cameras, as there is no requirement for sensors and observed subjects to face each other for the extraction of features. However, as for the visual modality, the difficulty when identifying speakers in an uncooperative, uncontrolled setting, is that far-field sensors need to be used, such that noise, cross-talk, room reverberations, low signal-to-noise ratios, etc., make the identification task incomparably harder. This is in comparison to using close talking or lapel microphones, which are worn by each room occupant. In the latter case, both segmentation and identification of a speaker's voice are easier to achieve. Usual state-of-the art techniques perform an analysis in the frequency domain, using features such as Mel Frequency Cepstral Coefficients (MFCCs) of Perceptual Linear Predictive (PLP) features in combination with Gaussian Mixture Model (GMM) classifiers [91].

In the case of far-field speaker recognition for multiple persons, more specialized techniques must be employed. The segmentation of speech itself becomes a larger issue. The task for voice identification becomes that of: 1) Determining segments of speech as opposed to noise, non-speech sounds or silence, 2)

dividing the speech segments into portions belonging to different speakers and 3) identifying the speakers in each segment. The complete process of detecting speaker turns and assigning speech segments to speakers is referred to as speaker diarization. Performing diarization in a smart environment (as opposed to telephone channels, for example) offers the advantage that information about the source of speech, derived through microphone array analysis, can help in the process of segmenting speech or, through beamforming, to enhance the quality of the audio signal used in identification. Several methods have been recently proposed to perform far-field diarization and identification, partly making use of microphone array localization [92; 13; 17; 88; 36; 69].

In [13], Ajmera et al. present a system that performs speaker segmentation and clustering without prior knowledge about the number of speakers or their identities. The algorithm works by first oversegmenting the audio data and training speaker models for each segment. MFCC as well as LPCC (Linear Predictive Cepstral Coding) features are extracted from each segment and used to train individual GMMs. The authors then cluster the trained models and segments using a modified form of the Bayesian Information Criterion (BIC), which automatically determines the optimal number of clusters without the need for any parameter tuning. They present results on the three benchmark datasets used by NIST (The U.S. National Institute of Standards and Technology) which show the effectiveness of the approach. The merit of the clustering approach is that it performs segmentation and speaker association jointly, on difficult and noisy data. The downside, however, is that it is not a run-on technique. It performs batch processing of entire audio segments and is therefore not suited for realtime systems. Another point is, of course, that the true identity of speakers is not inferred, i.e. since no comparison is made to pre-trained speaker models, only a separation of the data is achieved and no direct association to persons is made. In [12], they extend their approach to include also the information on the time delays of arrival of speech for microphone array pairs in a meeting situation. The TDoAs are included alongside the MFCC features in the agglomerative clustering process. The authors show that the localization information derived from microphone array analysis can substantially improve the diarization performance.

More recently, Pardo et al. in [88] propose a clustering method similar to that of Ajmera et al., combining acoustic MFCC-based features with TDoAs from multiple distant microphones to perform speaker diarization in meeting scenarios. Additionally, the authors use a beamforming technique based on computed TDoAs to enhance the speech signal before speech/non-speech segmentation and clustering is performed. Experiments were made on the NIST RT02-RT06 datasets (RT stands for Rich Transcription and is a series of evaluations hosted by NIST in the acoustic speech recognition and diarization domains). They show that beamforming, as well as the inclusion of localization information improve diarization performance. They also show that diarization based solely on

time delay information is also feasible, although results stay below those reached with fusion. For these experiments, the assumption is made that speaker locations do not change, as this would disrupt the clustering. This is, of course a limitation and the authors acknowledge that the general case could involve the tracking of exact speaker locations. This was not be done, though, as the information about relative microphone array positions was not available.

The above mentioned approaches realize the detection and segmentation of acoustic signals, and the association to speakers using multiple sources and without prior knowledge of the number of present speakers. However, they do not address the problem of actually identifying the actual speakers, i.e. determining which of a set of known persons, if any, the speech should be attributed to. The problem in the identification case is that a correct mapping of segments to speakers has to be realized e.g. *across recordings sessions*, though recording conditions, such as room characteristics, etc. may change. This is typically realized by training in a set of models (usually GMMs) for known speakers beforehand, which will be used to classify extracted speech segments, and applying normalization and warping techniques to alleviate the effects of changing signal characteristics. These may be observed in the far-field identification case due to differing sensors with differing speaker distances, changes in speaker locations, in room characteristics, such as reverberation, etc.

Barras et al., Luque et al. and Ekenel et al. in [17; 69; 36] present systems for the identification of speakers in a small meeting scenario using distributed microphone arrays. MFCC and PLP features, as well as modifications thereof were used in conjunction with GMM classifiers trained using Expectation Maximization. The experiments were made on presegmented audio segments of varying lenghth for 28 speakers in a closed set identification task. Although silence, noise, and a small amount of cross talk could be included in the individual segments, this means that diarization, i.e. the association of segments to speakers could be achieved simply by performing sequence level identification for whole segments using the speaker-dependent models. The results presented in [17] show that beamforming improves recognition accuracies, compared to single microphone analysis. The results from the other two approaches, however, show that a decision-level fusion of identification results from multiple microphones bring an even greater gain in performance. Finally, all approaches demonstrate that temporal fusion greatly improves identification performance, with accuracies increasing proportionally to the length of examined segments. One must note though, that this form of temporal fusion can be used most effectively in cases where speech segments can be unequivocally assigned to specific speakers (i.e. no diarization needs to be performed). In the opposite case, the use of lengthy speech segments for identification may actually lead to more errors, as the chances of accidentally including speech from multiple speakers in the same segment rise.

For both the acoustic and the visual modality, state-of-the-art approaches have shown that a temporal fusion of observations, and fusion of several sensory sources can lead to improved identification accuracies. In [101; 20; 68; 36], it is shown that this can also be true of the fusion across modalities, although this is not always the case. Especially when few observations are available, such that accuracies for single modalities stay relatively low, a decision-level fusion of audio-visual results, e.g. based on confidence weights as proposed in [101; 36], improves accuracies. In cases a single modality dominates the results, however, for example as the number of observations increases, multimodal fusion can lead to a degradation in performance. In [68], the problem is circumvented by predefining weighing factors for different observation lengths, based on enrollment data. This solution, however, is highly dependent on the data used, and cannot be applied in the general case.

Most of the work on multimodal fusion for identification is made in the domain of biometric identification or verification where, among others, the problems of detecting and extracting features, and associating observations to persons are often implicitly solved beforehand. In biometric verification, one can usually assume the cooperation of users, such that no data association is necessary and features from all modalities under consideration can be extracted simultaneously. One early example is the approach presented by Choudhury et al. in [28], where the identities of users operating an ATM machine are verified audio-visually. Some amount of research has been done in the domain of biometric identification to determine optimal criteria for multimodal fusion [59; 93; 53]. In [93; 53], the authors show that a decision level fusion of modality results, based on confidence scores boosts recognition performance in the general case. Scores for separate modalities are normalized using min-max normalization, Z-scoring, Tanh estimators, sigmoid functions, and so forth. The combination of scores using several methods, such as the sum rule, product rule, maximum rule, etc., is investigated for the open set identification case. The authors note that even after normalization, the scores for genuine and impostor classes are not normally distributed. In fact, the distributions may differ significantly from one modality to the next. Due to this, score-based likelihoods should not be used directly in modality fusion. They propose a Parzen-window estimation method to convert matching scores into posterior probabilities for the genuine and impostor classes, based on training data. The resulting probabilities are values in the range $[0, 1]$ and can be treated as normalized (or warped) scores for fusion. The authors note that the min-max, Z-score, and Tanh normalization techniques followed by the sum rule result in the best recognition performance.

Another example of multimodal fusion using score normalization and confidence weighting is presented in [43]. Here, the confidence scores are used both for temporal and for modality weighting in an open set identification scenario. The scenario involves single users engaged in a dialog with a humanoid robot system, and the identification is made on sequences of observations using face and

voice identification. Here again, min-max score normalization is applied and confidence values are computed by considering the n-best list of scores for the highest ranking candidate identities. The authors investigate several methods for the estimation of confidence values, including the distance between the 2 best scores for single ID trials, the agreement of best hypotheses through time for separate modalities, etc. Their results show that confidence based scoring, both temporal and multimodal, significantly improves the overall recognition rate, and that confidence values can further be used to realize open set identification using classifiers trained only for the closed set case. As with most other multimodal fusion approaches, though, the presented system is designed to operate in controlled conditions, in a restrictive setup involving at most one user and requiring explicit cooperation or interaction (such as the use of a dedicated microphone, facing the robot head, etc).

2.3 Identity Tracking

As presented in the previous sections, a large amount of research was made on the individual components usable for open set multiple identity tracking: Visual, acoustic and audio-visual tracking, speaker diarization, face recognition, voice recognition, multimodal fusion of classifiers, etc. Not much work was done so far, though, on methods that combine *all* of these components in order to realize the open set identification and tracking of multiple users in unconstrained environments. This section now presents approaches that follow this higher level goal and solve the associated problems to some extent.

One of the earlier attempts is presented by Yang et al. in [116]. The authors present individual components for person and face tracking, face identification and voice identification and discuss a framework for their integration with the goal of realizing automatic meeting summaries. The integration is, however, still made conceptually on a frame level, assuming most cues for fusion are accessible at every point in time, and does not consider the multiple problems posed by data association, temporal and multimodal fusion. Further, the observing sensors are assumed placed on the meeting table in a frontal setup, which restricts the freedom of users. Other approaches that propose conceptual frameworks for global tracking and identification are presented e.g. by Rudnicky et al. [94] and Stanford et al. [99]. More recently, Menon et al., in [76], present an identification framework for smart indoor environments, based on tracking and face and voice identification. Only an abstract framework is presented, though, without actual technical realization or evaluation. The interesting point about their approach is that the tracking of identities is to be realized based on zones in a wide, distributed smart space comprised of several rooms. This is in contrast to most smart environment research involving audio-visual processing, where

the tracking attempts to estimate actual 3D locations in more restrictive, single room setups.

Recognizing the difficulties of face recognition using distant cameras, Hampapur et al. [44] perform 2D and 3D blob tracking on images from two fixed cameras and steer pan-tilt-zoomable (PTZ) cameras to zoom in on faces. They locate head regions by analyzing the silhouettes of tracked persons and discuss several strategies for target selection and active camera assignment to capture good facial views. They do not, however, address the problem of recognizing users in the closeup views or of fusing identification results over time.

Similarly, Stillman et al. [107] utilize visual tracking in fixed camera views and acoustic source localization from microphone arrays to steer active cameras and perform face detection and recognition of multiple users. They do not, however, offer a framework for fusion of the identification results or for identity tracking through time.

You et al., in [118] present a smart interaction environment where multiple users interact with large displays. A fixed view allows the tracking of multiple users in front of the display. Frontal views of faces are automatically identified and associated to user tracks. The authors present no method for identification confidence estimation or temporal fusion, though. Also, they rely on the flawless identification results to adapt color models for tracking and on flawless tracking results to keep identities associated to persons in time. In realistic, natural scenarios, however, both these prerequisites can seldom be met.

In [50], Huang et al. present a smart room setup, consisting of a combination of an omnidirectional camera and distributed arrays of PTZ (pan-tilt-zoom) cameras. Persons are tracked using a simple foreground blob analysis in the view of the omnidirectional camera placed on the table, PTZ cameras are steered at person locations, faces in individual views are tracked and identified and score-based temporal fusion is performed. Identification is only performed, though, for face tracks in single camera views, and identities are not kept as the views change. Also, no framework for data association and global multi-person identification is presented.

In [77], Mikic et al. present an identity tracking system for smart meeting rooms. It performs blob-based 3D tracking of room occupants in fixed camera views and triangulates 3D person positions. It also performs face identification and voice identification for person entering the space. When new persons are detected by the tracking algorithm, the best view for observation of the face is determined and a snapshot is taken. The user is also required to speak at the time of face capture such that a combination of face and voice identification results can be performed. The identities of persons are kept through continuous tracking in the smart space, with subsequent speech identifications serving only to determine the active speaker. While the system combines tracking and

multimodal identification, it does not perform temporal fusion of continuous observations and cannot recover from tracking mistakes. Also, constraints in the way users behave when entering the smart space (facing the direction of walk, speaking) reduce the naturalness of the scenario and limit the applicability of the system.

In [22], Busso et al. present a tracking and identification framework for a similar scenario. Their approach uses several corner cameras and an omnidirectional camera on the table to track persons and faces. A 16-channel microphone array is also used for source localization, beamforming and speaker identification. While the system keeps continuous visual track of users, it does not, however keep track of their identities, as the audio and visual information is used only to determine the identity of the currently active speaker. The result is a speaker diarization system, which incorporates visual information to improve accuracies and infer exact person locations. Temporal fusion of identification results associated to person tracks, confidence estimation or multimodal fusion are not performed.

Similarly to the approach by Mikic et al, but much more recently, Salah et al., in [95] present a system for the tracking and identification of multiple users in a smart environment. They perform audio-visual particle filter based person tracking, source localization, speech segmentation and identification, and face identification in close up views of the entrance door. They evaluate their method on two recorded audio-visual sequences and show the advantages of multimodal fusion for keeping track of identified persons in time, compared to visual analysis alone. The approach does not, however, perform temporal fusion for acoustic or visual identification, nor confidence-based multimodal fusion. Facial identification is performed for each occupant using a single captured face shot at the room entrance. Similar to [77; 118], the system relies on the perfect identification of users upon entrance to build person specific appearance models for tracking, and relies on flawless tracking to keep correct identities associated to users in time.

All the approaches presented above tackle the problem of tracking and identifying multiple users in smart environments. While many of the necessary components are realized to some degree in each approach, none them jointly handle all the associated problems of audio-visual tracking, multimodal identification, temporal score-based fusion, source localization, data association and open set identification necessary to reliably recognize known persons and ensure continuous and robust tracking of identities. Additionally, all the approaches that attempt a continuous identification of multiple users rely on an accurate detection and tracking of all users and perform identification based on the results of the tracking step. If the tracking component fails, or mistakes are made, the accuracy of the overall system degrades and, often, errors can not be detected and recovered from. As shown in the previous section, though, the

flawless tracking of all persons in realistic, cluttered smart environments, and the continuous availability of observations are not realistic prerequisites. For the development of truly robust systems, failures of individual components should be considered in the system design and handled efficiently and automatically, without the need for user intervention or manual tuning.

In this thesis, a method is proposed that integrates all the above described components in a joint probabilistic way. Persons are automatically detected and tracked at any location in the smart space without the need for cooperation or interaction. Person models are built on-the-fly and continuously adapted during the whole observation sequence. Person tracks are estimated audio-visually using a variety of features extracted from multiple cameras and microphone arrays in a robust, probabilistic filtering framework. Face and voice observations are automatically captured and probabilistically associated to persons, even in the presence of severe occlusion, low sensor coverage or missing person tracks. Identities are derived jointly for all persons using the sequence of associated observations. Identification is made using temporal and cross-modal fusion, based on the combination of normalized confidence scores, and for the *open set* identification case. The method is extensively evaluated on a large benchmark database of realistic recordings using well defined metrics and a systematic evaluation procedure. Quantitative results allowing to judge the influence of various components, or failures thereof, on overall tracking and identification performance are presented.

The evaluation metrics are also defined in the course of this thesis, which is why an overview of the state-of-the art in the evaluation domain is given in the following section.

2.4 Performance Evaluation

In recent years, there has been a growing interest in performing systematic evaluations of tracking approaches with common databases and metrics. Examples are the CHIL [3] and AMI [1] projects, funded by the EU, the U.S. VACE project [9], the French ETISEO [6] project, the U.K. Home Office *iLIDS* project [7], the CAVIAR [2] and CREDS [122] projects, and a growing number of workshops, such as e.g. PETS [8], EEMCV [5], and more recently CLEAR [4], to name just a few. The problem faced by all major evaluation efforts where more elaborate multiple-target trackers are evaluated is the definition of suitable metrics for quantitative measurements and comparative benchmarking. The trend observed is that new evaluation projects also bring along their own newly created multi-target tracking evaluation metrics. Although well defined, commonly agreed on metrics exist for single object trackers, making benchmarking rather straightforward, there is still no general agreement on a principled evaluation

procedure using a common set of objective and intuitive metrics for measuring the performance of multiple object trackers.

Li et al. in [64] investigate the problem of evaluating systems for the tracking of football players from multiple camera images. Annotated ground truth for a set of visible players is compared to the tracker output and 3 measures are introduced to evaluate the spatial and temporal accuracy of the result. Two of the measures, however, are rather specific to the football tracking problem, and the more general measure, the "identity tracking performance", does not consider some of the basic types of errors made by multiple target trackers, such as false positive tracks or localization errors in terms of distance or overlap. This limits the application of the presented metric to specific types of trackers or scenarios.

Nghiem et al. in [84] present a more general framework for evaluation, which covers the requirements of a broad range of visual tracking tasks. The presented metrics aim at allowing systematic performance analysis using large amounts of benchmark data. However, a high number of different metrics (8 in total) are presented to evaluate object detection, localization and tracking performance, with many dependencies between separate metrics, such that one metric can often only be interpreted in combination with one or more others. This is for example the case for the "tracking time" and "object ID persistence/confusion" metrics. Further, many of the proposed metrics are still designed with purely visual tracking tasks in mind.

Smith et al., in [97], also attempt to define an objective procedure to measure multiple object tracker performance. However, a large number of metrics is introduced: 5 for measuring object configuration errors, and 4 for measuring inconsistencies in object labeling over time. This makes it hard to get a clear overview of overall tracking performance [117]. Some of the measures are defined in a dual way for trackers and for objects (e.g. the MT/MO, FIT/FIO, TP/OP). This makes it difficult to gain a clear and direct understanding of the tracker's overall performance. Moreover, under certain conditions, some of these measures can behave in a non-intuitive fashion (such as the CD, as the authors state, or the $\overline{FP}$ and $\overline{FN}$, as will be demonstrated in Chapter 5).

To remedy the lack of clear and intuitive metrics, a thorough procedure was developed, allowing to detect the basic types of errors produced by multiple object trackers and two novel tracking metrics are introduced, the Multiple Object Tracking Precision ($MOTP$), and the Multiple Object Tracking Accuracy($MOTA$), that intuitively express a tracker's overall strengths and are suitable for use in general performance evaluations.

In addition to providing a novel theoretical framework, some of the proposed metrics have been used in two international evaluation workshops, which can be seen as field tests for their applicability. These evaluation workshops, the

CLEAR – CLassification of Events, Activities and Relationships – workshops, featured a variety of tracking tasks, including visual 3D person tracking using multiple camera views, 2D face tracking, 2D person and vehicle tracking, acoustic speaker tracking using microphone arrays, and even audio-visual person tracking. For all these tracking tasks, each with its own specificities and requirements, the here introduced $MOTP$ and $MOTA$ metrics, or slight variants thereof, were employed [102; 103]. The MOT metrics are also employed in a growing number of publications outside of CLEAR [47; 65; 23; 95; 56].

Additionally, a novel measure for the evaluation of *identity tracking* performance was introduced. While the tracking and localization of identities has been attempted for quite some time in various approaches, as presented in the previous section, the evaluation of performance was up to now realized in a rather ad hoc way, without clear and well defined metrics that account for all relevant error factors. To remedy this, a definition of the general *identity tracking* task is introduced in Chapter 5, a formal evaluation procedure is defined and a novel metric for quantitative evaluations is presented. This metric, the Multiple Identity Tracking Accuracy ($MITA$), expresses a trackers ability at localizing *and* identifying known persons in a larger group, for the open set scenario.

3 Features for Identity Tracking

In this chapter, the various features used for unobtrusive multiple identity tracking are presented. They include both low-level and high-level features for person detection, tracking, speaker localization, face recognition and voice identification. In this context, low-level features represent those which can be more or less directly extracted from the sensor streams, whereas high-level features are the results of more complex analysis and express some form of spatio-temporal relationship or carry a symbolic meaning. These are usually the outputs of more or less complex classifiers or trackers operating on the lower level features. Examples for high-level features are detection windows from image-based person detectors, computed 3-dimensional sound source locations, or hypothesized identities resulting from facial identification.

There are plenty of reasons speaking for the use of a rich variety of features. In some cases, different features work concurrently to help obtain certain types of information, in other cases, they work in a complementary fashion:

- As mentioned before, features allowing the identification of faces or voices may be unobservable for lengthy periods of time. Using both types may help speeding up the identification of a person. Similarly, it may be hard to detect persons in the scene using certain camera views or detection methods. A combination of detection techniques, based on appearance patterns, motion, foreground segmentation, etc., is necessary to achieve better results.
- As a continuous identification of all occupants is not feasible, it is necessary to keep track of already identified persons until features for identification can again be observed. The features used for tracking are generally much simpler than those used for identification. They nevertheless help bridging the gap in higher-level information.
- Single observations about a person's identity may be inaccurate or noisy, due to various reasons. For example, face ID accuracies may be influenced by head pose, lighting or resolution, etc., whereas speech identification may generate faulty hypotheses due to cross-talk or unknown sound sources. A way to boost recognition performance is to accumulate all observations made for a specific person and perform temporal filtering or fusion. It has been shown that temporal fusion can significantly increase

the performance of person identification techniques under challenging conditions (just one example are the CLEAR Person Identification evaluations [106; 105]). In a multiple user scenario, however, observations need first to be associated to the correct person for accumulation and temporal fusion to be possible. As the correspondence in our case is not given beforehand, the data association problem is a non-negligible factor affecting performance. This is another example of where tracking features need to be used concurrently with identification features.

- Even when persons cannot be accurately detected or tracked, their presence in the smart space may still be correctly inferred from identification features. This may be the case when the sensors used for identification differ from those used for tracking (e.g. table top microphones and possibly steerable cameras with limited field of view, as opposed to microphone arrays or wide-angle cameras overlooking the space). An example scenario would be one in which 2 voices are recognized in the smart space with high confidence, while only one persons is tracked and correctly identified. The logical consequence would be that the second person is present, although his or her location is unknown. In the context of identity tracking, this information, although less detailed, is undoubtedly also useful.
- Finally, in the proposed framework, new persons entering the environment are tracked and data association is performed using automatically learned in person models that are themselves continuously refined as new observations come in. In this context, building on a rich set of varied features is extremely useful. It allows to detect consensus among feature types and thereby reject faulty observations or associations as outliers.

For extracting the various lower and higher level features, many different algorithms, classifiers, trackers and preprocessing steps may be used. The focus in this work is not on the concrete techniques or implementations used. Rather, state-of-the-art techniques are employed whenever possible and the focus is put on the way these need to be adjusted and fused to serve as inputs to the identity tracking. The idea is that the ID tracking framework should be generally applicable using a wide variety of perceptual components, without the need for specialized techniques designed for specific application scenarios. In the following, the techniques used for extraction of the different feature types are only briefly described, focusing on their algorithmic strengths and weaknesses, and references to related work, giving detailed explanations, are made when appropriate.

3.1 Overview of Feature Types

The identity tracking features used in this work can be roughly separated into visual features and acoustic features.

On the visual side, these are:

- Foreground support maps (foreground feature). These are computed in the individual camera images and serve as rough hints as to where persons may or may not be.
- Person detection windows, in the following briefly called person detections (detection feature). They are the outputs of detectors working on the views of cameras installed in the corners of the observed space. Specifically, detectors for the upper torso of persons are used here.
- RGB color responses (color feature). RGB color models of the upper torso of tracked persons are built, based on the output of person detectors, and used to search for specific persons in the camera images.
- Foreground blob tracks in top-view images (top view feature or "top feature"). They are the output of a simple tracker operating on the foreground support maps from a ceiling mounted panoramic camera. The reasons for using such types of features will be explained later on.
- Localized, identified faces in the scene (face ID features). These are the output of a chain of processing steps performed on the views of corner cameras: Face detection, 3D-localization and identification.

On the acoustic side, features are:

- Identified voices (speaker ID features). These are obtained by analyzing the signal from an omni-directional microphone placed in the smart space. They are the results of speech detection, segmentation and identification. They are usually accompanied by
- Speech source locations in the scene. These are computed using a number of microphone arrays placed at the edges of the smart space. When speech localization was successful, we speak of localized speaker ID features; if not, "non-localized" speaker ID features are obtained.

The various audio-visual features have quite different characteristics that prohibit their straightforward combination, e.g. in a synchronized, feature-level fusion scheme. A more detailed categorization reveals strengths and weaknesses of different feature types in several dimensions:

- Discriminative power (specificity to a given person). Face ID and speaker ID features have the strongest discriminative power as they are in principle specific to a unique person. Features based on the color of clothing are less specific, as they can be shared by several individuals, yet in many scenarios with a closed set of persons, they can be very helpful in resolving ambiguities. Finally, foreground or detection features usually have no discriminative property.
- Detection accuracy (specificity to persons in general as opposed to other objects). This characteristic pertains to the usefulness of features when it comes to detecting persons in the first place. Here, ID features and detections rank highest. Foreground maps or blob features rank lower, as they can be caused by shadows, changes in lighting, moved objects, etc. Color features, finally, can not serve to detect the presence of persons in general.
- Observability. This characteristic pertains to the frequency or regularity, with which features can be extracted. Identification features, with high accuracy and discriminative power, can usually only be extracted on an irregular basis. Faces are only identifiable when they are turned toward a camera. Speaker ID features rank very low, as they are only available when a person speaks, but also torso detections may, depending on the scenario or the overall body pose, not always be available. This is in strong contrast to foreground or color features which, save for occlusions, are observable at video frame rate whenever the concerned person is in the field of view of the sensor.
- Spatial accuracy. Some features, such as detections, localized speech or blob tracks can be directly associated with a given 3D location or line of view in the scene. Others, such as foreground or color features are presented in the form of support maps and can only indirectly be used to infer person positions. The former are extracted and input to the tracking framework in a bottom-up fashion, while the latter are used in a top-down fashion, to verify hypothesized person locations. Non-localized speaker ID features, finally, only give an indication of the presence of a person, without any indication of location.
- Initialization. As opposed to foreground and person detection features, colors and audio-visual ID features can only be used if specific models of the persons to be recognized exist. Color models need to be initialized and updated on-the-fly, based on observations which can only be made once a person has been detected with high enough accuracy. Face and voice models, on the other hand, may be trained in beforehand for a set of known persons. Nevertheless, the identities of the persons actually present in the scene still needs to be inferred from noisy observations before the so obtained reduced set of models can be used e.g. in data association.

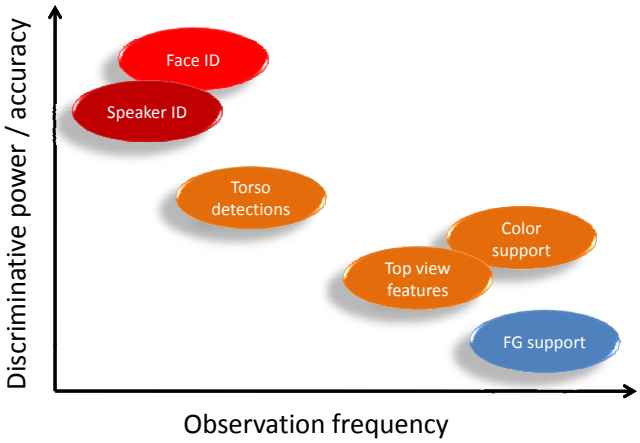


Figure 3.1: Overview of features for identity tracking. Foreground and color support maps can in principle be observed at framerate. They are not reliable enough, though, to accurately detect the presence of persons or distinguish them. Face and speaker ID cues are the most discriminative and most accurate features. They are, however, much harder to observe using distantly placed sensors. No single feature combines all the advantages of high observability, high accuracy and high discriminative power, such that a combination of features should be used.

- Constancy. Certain aspects of features, such as those pertaining to identity, are expected to remain relatively constant in time, whereas others, such as location, may vary considerably from one moment to the next.

Figure 3.1 shows a coarse qualitative categorization of the various feature types from the points of view of discriminative power, accuracy and observability. All the mentioned features are fused in the identity tracking framework. The following classification into tracking and identification features is therefore by no means strict (features used for identification can also be used for detection, which in turn helps tracking). It is motivated by the difference in feature complexity and observability, with simpler tracking features bridging the gaps in observability for higher-level identification features, and serves only as a rough functional categorization.

3.2 Features for Tracking

3.2.1 Foreground features

Foreground support maps are commonly used in tracking applications involving fixed cameras. In our case, these maps are obtained on gray-scale images, using a simple procedure: First, a background model is obtained by pixel-wise averaging of a few initialization images. These images depict the smart environment, preferably without occupants. Then for each time frame, the background model is subtracted from the input image and the absolute pixel differences are thresholded using a global fixed threshold T_{FG}, resulting in the final support map. To accommodate for small changes (e.g. due to lighting changes, shadows, or small moved objects), the background model is continuously adapted with each input frame using a small learning factor α_{FG}.

Although much more sophisticated foreground segmentation techniques exist, for example using individual pixel thresholds per pixel, Gaussian mixture models for adaptation [100], etc., these were not implemented. Even using state-of-the-art techniques, the obtention of a clean, well segmented foreground map, depicting only person regions, is only possible in controlled, uncluttered environments, by building on strong assumptions about the movement of objects of interest with respect to the background. As these assumptions can not be guaranteed in our case, the foreground feature can only serve as a coarse indicator anyway, which is why the simple method described above largely suffices for our needs.

In the case of the CLEAR seminar recordings, a set of background images, depicting the empty smart environment, is provided for most of the evaluation sequences. Whenever available, these images are used for initialization of the background model. In the contrary case, it is initialized using the first few frames of the recording. As persons are often already present in the space at the start of recording, this of course reduces the quality of the background model, and consequently of the obtained foreground maps.

For fast computation of the foreground support in a given subwindow of the input image, the integral image method [113] is applied. In an integral image, each pixel represents the sum of all pixels above and to the left of it in the original image. It can be precomputed efficiently in one pass for the whole image and later allows to quickly calculate the sum of pixel values in a subwindow of arbitrary size, using only four table lookups. This is a prerequisite for the efficient computation of observation models in the subsequent particle filter framework, as explained in Chapter 4. An example foreground map for one of the camera view of a CLEAR seminar is shown in Fig. 3.2.

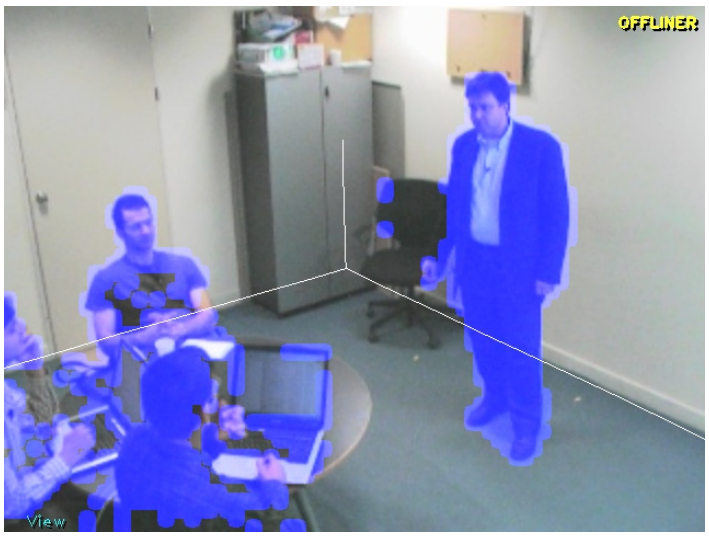


Figure 3.2: Example foreground support map in a camera view of a CLEAR Seminar recording. Clean support maps allow for reliable tracking. Shadows and changes in the environment often perturb the foreground support, though.

3.2.2 Detection features

In the context of this work, the term "detection feature" is used as an abbreviation to designate the output of an appearance-based person detection process in a camera view. More specifically, the upper torsos of persons in the smart space are detected and the resulting detection windows used as inputs to the tracking process. The choice of the upper torso as the region to detect can be motivated as follows: Firstly, it is observable in most common situations, as opposed to e.g. the legs, which are frequently occluded by objects, chairs, or when persons are sitting at tables, etc. Secondly, it is large enough to be detected reliably in the views of distantly placed cameras, as opposed to faces, e.g., for which low resolution and relative orientation still pose a greater challenge.

Accurate person detection is a basic requisite in tracking tasks involving automatic initialization. Only if we can determine the presence and location of a person with high enough confidence can we construct person-specific models and create new person tracks. If the decision is based on unreliable features, such as motion or foreground segments, which may be caused by shadows or other objects, we run the danger of learning in faulty models of the background, chairs, etc., resulting in persistent false tracks. The vast majority of state-of-the-art appearance-based person detectors in the literature build on a more or less efficient computation of gradient descriptors in intensity images, and the automatic training of robust classifiers using large amounts of training data

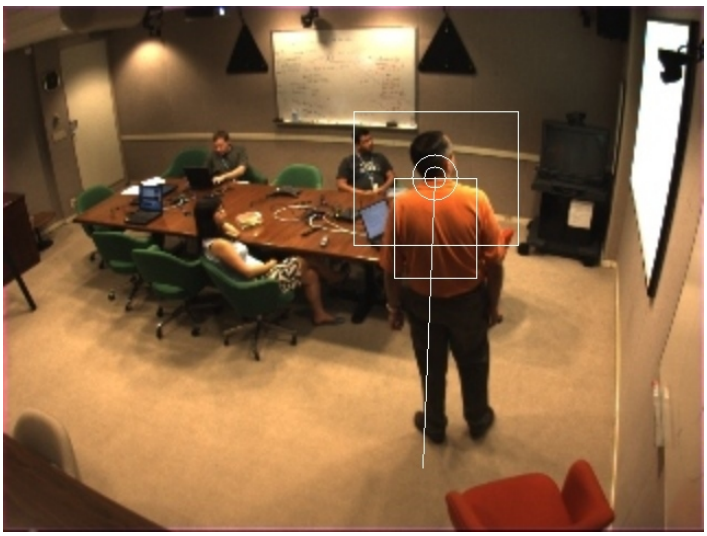


Figure 3.3: An upper torso detection in a corner camera view. The larger bounding box represents the actual window returned by the detector. Colors for adaptation of person specific models are extracted from a much smaller window, though, to avoid mistakenly including background colors.

[113; 31; 18]. They can usually be adapted to detect any kind of pattern with relatively stable appearance.

The detectors used here are those introduced by Viola and Jones in [113; 66], and initially proposed for the task of face detection. They are based on rectangular, box-like approximations of Haar-Wavelet descriptors, which are efficiently computed using integral images. The detection is performed by shifting a window of variable size on the input image and classifying the so defined image regions. The efficient selection of features and training of classifiers is made using a boosting technique, Adaboost, and the fast classification required by the exhaustive scanning step is achieved by organizing weak classifiers into cascades of increasing complexity. Figure 3.3 shows an example bounding box resulting from an upper torso detection.

In principle, any other realtime-capable detection algorithm, for example the histogram of gradients (HOG) approach [31], based on Support Vector Machine (SVM) classification, could be employed at this stage. The result of detection, in any case, is for each frame a list of bounding boxes representing hypothesized person locations.

The main difficulty in torso detection is caused by the variable appearance of upper bodies in the scene. Common detectors are designed to recognize standardized poses, such as frontal torsos, faces, etc., with the detection of multiple orientation classes usually associated with a relatively high computational over-

head. The problem is made worse in natural, unconstrained scenarios by the posture of observed persons (for ex. leaning back strongly in a chair, bending over a table, resting on an elbow), by partial occlusion (for ex. in some views by the backs of chairs). The problem may be alleviated to some degree by the use of several cameras observing from different angles. Nevertheless, in many situations, one cannot rely on the frequency of detections to ensure accurate tracking.

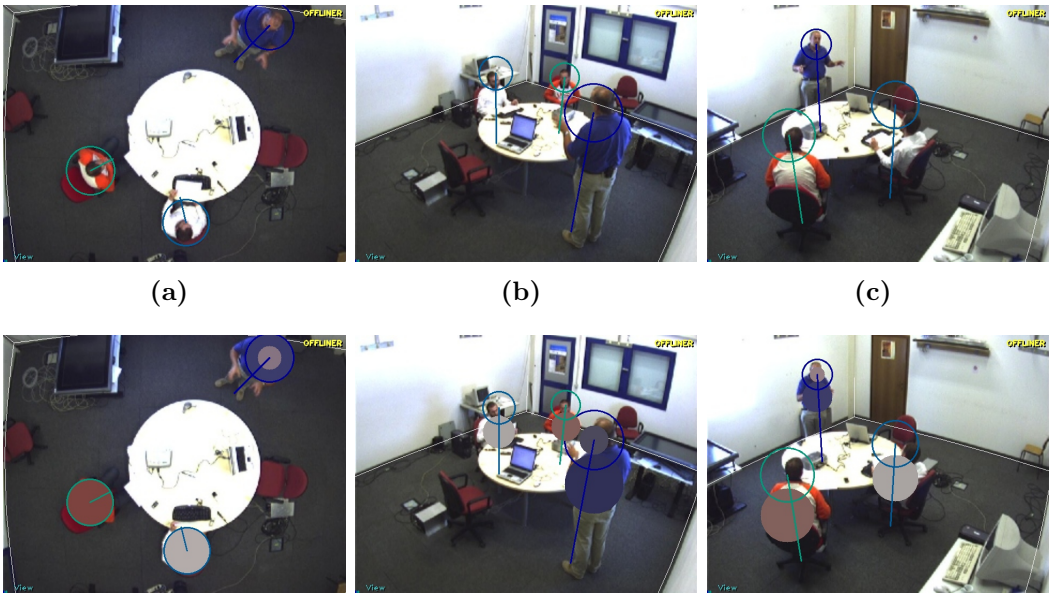
3.2.3 Color features

Colors are widely used for tracking tasks in general as they offer an additional source of information to discriminate between between objects of interest. They offer the advantage that the object or person of interest can be easily detected and tracked as long as its representative colors or "texture" is sufficiently distinct from others and the background. The downside to using texture or color models is that they need to be learned in either beforehand, in a separate initialization step [62], or on-the-fly for each new initialized track. Further, the appearance of colors may vary considerably from one camera sensor to the next, or even through time for one same sensor, as it is strongly influenced by ambient or local changes in lighting. Here, color features are used to describe the upper torso region of subjects in the smart space. The reason for limiting the color model to the upper torso is the same as mentioned above for detections: It is the largest, best visible region and is least subject to occlusion by tables, chairs, other persons, etc. To automatically learn in color models for a tracked person, the person region first has to be segmented more or less accurately from the background. It has been shown that the quality of segmentation can greatly influence the quality of resulting color models [15]. The approach chosen here to address this problem is to extract color features only inside the bounding boxes delivered by the detection step from Section 3.2.2. The chosen representation is quite simple, and consists of the mean of RGB color values inside the detected upper torso region, as well as the standard deviation of mean values accumulated in time. So for a single detection region R_{det}in one view, the extracted color feature

$$col_{det} = \frac{\sum_{p \in R_{det}} col_p}{|R_{det}|}$$

with $col_p = (r_p, g_p, b_p)$, the tuple of RGB values for pixel p and $|R_{det}|$ the number of pixels in R_{det}. The color model of an upper torso for one view then consists of $\{\mu_{col_{det}}, \sigma_{col_{det}}\}$ with

$$\mu_{col_{det}} = \frac{1}{n} \sum_{i=1}^{n} col_{det_i}$$



(a) (b) (c)

(d) (e) (f)

Figure 3.4: An example of learned upper body color models for multiple subjects. The top row represents the original views from three cameras. The colored circles represent person tracks. The bottom row shows the mean color models learned individually in each view for each person in an unsupervised way.

$$\sigma_{col_{det}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(col_{det_i})^2 - \mu^2_{col_{det}}}$$

the mean and standard deviation of color values accumulated for the same torso in time.

To account for the sometimes large differences in apparent color from one camera to the next, a separate color model is built for each view using only detections from that view. Colors models for views where no detection could be made are computed as the average of color models from other views. As soon as specific observations for a view become available, though, an individual color model is again used. Figure 3.4 shows an example of upper torso color models learned in for multiple tracks in several camera views.

The choice of the mean RGB color as color representation may be seen as simplistic or inaccurate in comparison to other, much more complex representations, such as color histograms or the MPEG7 dominant color descriptor [15]. Yet, there are two reasons speaking for this representation in our case: Firstly, the mean color is less affected by small changes which often occur when a person

moves or changes its relative orientation to the camera. Secondly, the mean color can be computed very efficiently for any subwindow of the image, as opposed to, for ex., color histograms, by using integral images. The evaluation of a simple Gaussian color model, represented by its mean and standard deviation, is also much faster than standard techniques for histogram matching.

The color feature is used in two ways:

- Support maps for the color feature are built in the form of individual integral images for the R, G and B channels. These support maps are later used in the tracking framework for the quick computation of mean color values inside hypothesized person regions, which are compared to the person models tracks.
- A mean color is computed for each detected upper body in a camera image. The representation of the detection feature is therefore extended with the color description: $feature_{det} = \{bbox_{det}, col_{det}\}$. This color information can later be used to learn in and adapt view-dependent color models for tracks, but also in a data association step, when mapping observations to tracks in the first place.

Additionally, the color feature is used in a consistency check, to filter out faulty detections. Owing to the nature of the detection algorithm, which is based on local image gradients, these sometimes occur in areas of low contrast, on plane surfaces, etc. This kind of error is difficult to eliminate in the detection step itself, as a reduction of the amount of false positives is accompanied by a reduction of the sometimes already quite low recall rate. In this case, color constitutes a separate source of information which can help reject erroneous detections. The idea is that a color representation is only useful for tracking a specific object if it is sufficiently different from that of the surrounding background. The filtering is accomplished for every detected torso by constructing a "background" bounding box, centered at the original detection and with three times the area. Again making use of the color channel integral images, a mean color is computed exclusively for the outside region and compared to the color inside the detection box. Only if a sufficient difference exists (measured by the mean and variance of RGB values inside the detection window) is the detection feature considered valid and passed on to the tracking step.

3.2.4 Top View Features

These features are the outputs of a simple blob tracker operating in the images of a top-view panoramic camera mounted under the ceiling of the smart space. The reasons for using this type of feature in top camera views, as opposed to other views, are as follows:

- Panoramic views from a top perspective offer the advantage that persons are much less likely to be occluded by each other or environmental objects. For views from wall-mounted cameras, with relatively low downtilt, person blobs resulting from foreground segmentation often merge when persons pass behind each other. They are also often cut off by partially occluding objects. All these factors make an analysis on such views much more complex. This is the reason why top views are used in a wide range of practical tracking scenarios involving multiple persons, vehicles etc. [2; 47]
- It is much easier to constrain the reasonable range of person region sizes in a top view, as distances to the camera are much less subject to change. This helps distinguish individual persons even in the case adjacent blobs should merge.
- Standard appearance-based person detection techniques, such as the one presented in Section 3.2.2 are difficult to apply in the case of arbitrary rotation of the target pattern. As opposed to wall-mounted, upright views, the assumption of a dominant (vertical) orientation of torsos can not be made. From a top perspective, persons may be observed as facing in any direction. Using a detector for multiple orientations usually comes at the cost of additional computation time or reduced accuracy. This is why a simpler method was adopted here, to serve as a "weak" detector operating on top view. For the purposes of this work, this method proved to be quite sufficient.

The top view foreground blob tracking algorithm works as follows: On the foreground maps described in Section 3.2.1, morphological filtering and a connected component analysis are performed to extract a number of possibly fragmented foreground blobs. These blobs are then mapped to simple circular person models in a local image-based tracking scheme. A person model in this case is represented by a center coordinate and a radius in the image. The image radius is approximated using an assumed average person torso width w_{torso}, the camera focal length and its height above the ground. Extracted blobs are associated to person models based on coverage in an Expectation-Maximization fashion: First, all blobs that are covered by a circular person model are associated to this model. Second, the model center is updated as the average location of all foreground pixels of associated blobs. New person models are initialized for each uncovered blob, deleted as soon as their blob support disappears, and fused when the distance between their centers falls below a certain threshold. The details of the algorithm, though of great practical value, are of limited scientific relevance and are therefore omitted here. Blob tracking has a long history in computer vision and a plethora of specific implementations for a great range of applications, all with their advantages and drawbacks, has already been proposed. The implementation used here should be seen as one of many means to

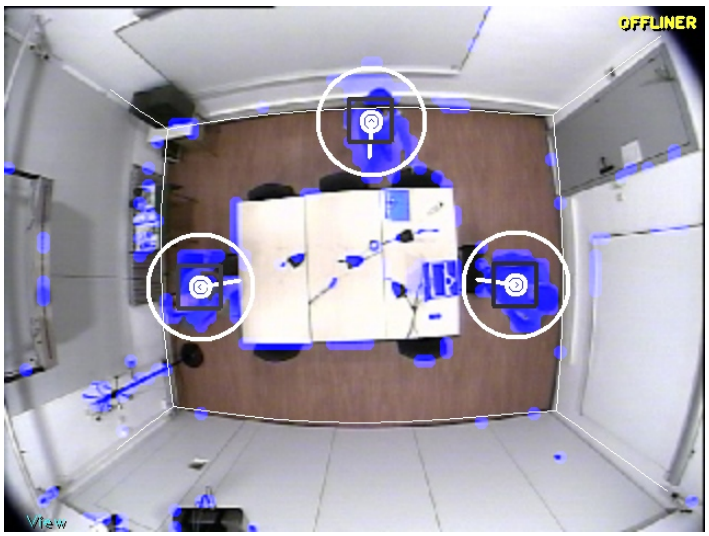


Figure 3.5: Examples of features extracted from a top view overlooking the smart space. The advantage of top views is that they mostly offer unoccluded views of the space and its visitors.

extract the type of feature we are interested in. Figure 3.5 shows an example output of the top camera blob tracking stage.

The advantage of features gained from top views has been demonstrated, for ex., in the CLEAR 2006 multiple person tracking evaluations [102]. In these multiple camera scenarios, systems that could make efficient use of the top view information outperformed sometimes more complex algorithms that relied only on other camera views. Although a fusion of all available views increases performance (see [4]), the top view can still be seen as the most important individual view.

As for upper torso detections, a color description is extracted from the area defined by the top view track and added to the top feature description. The person model's image center and radius are used to construct a bounding box from which the mean color is extracted as described in Section 3.2.3. As for detection features, top view features are filtered by comparing the mean color of the area inside the bounding box with that of the immediately surrounding area. For top view features, this proved especially useful to eliminate faulty tracks caused by small shadows, etc. The use of a bounding box for segmenting the torso region is of course a very coarse approximation. Because of this, and owing to the low accuracy of the foreground blob analysis (compared to the appearance-based detections from Section 3.2.2), a moderate portion of the background is usually included in the mean color computation, which degrades the quality of the obtained color models. As a consequence, the color description associated with top view features is less reliable than that of detection features.

The advantage of using bounding boxes is that again, integral images can be used for an efficient computation.

As a final remark, one should note that the top view features (or "top features") described in this section are not specific solutions tailored to the sensor setup being investigated here, to small rectangular rooms, etc. Rather, the use of top cameras in a tracking framework is common practice, and one can imagine extracting top features from a network of ceiling-mounted cameras covering a much larger space than that considered in the evaluations made here.

3.3 Features for Identification

The features considered in this thesis are those gained from face and voice identification. Consequently, the algorithms used for the recognition of faces detected in camera images and for the analysis of speech signals captured by far-field microphones are explained in this section. In the following, a brief explanation is first given, though, of the identification task itself and of the used evaluation measures, as these are relevant for understanding the design choices made in the rest of the section.

3.3.1 Classical Identification Tasks and Metrics

There are two classical tasks in person identification, both for the acoustic and visual modality: Closed set identification and open set identification.

Closed set identification refers to the task of recognizing a subset of persons from a closed set of known persons. Only persons which have been previously trained in are to be recognized and the only possible error is to confuse a known person with another. The accuracy measure for this kind of recognition task is the recognition rate, or correct classification rate (or simply recognition accuracy), which is given by:

$$ACC = \frac{N_{cc}}{N_{tot}}$$

with N_{cc} the number of correctly classified persons and N_{tot} the total number of persons to be classified.

Open set identification is, in comparison, a much more difficult task. Here, the objective is to recognize a subset S_{known} from a set $S_{tot_{known}}$ of known persons in a greater set S_{tot}, containing also unknown persons. The challenge here is that a number of persons which have never been seen before have to be classified as well. The recognition algorithm must therefore for each test person decide

if this person is known, and in the positive case determine his or her identity from the set of known identities. The subset of known persons to be recognized, S_{known}, is also referred to as "genuines", and the set of unknown persons in S_{tot}, also called "impostors", is denoted here by S_{unkn}. For each test person, the recognition system can make one of the following mistakes:

- Wrongfully determining the identity for a known person. This is referred to as "false classification".
- Rejecting a known person as unknown. This is called a "false rejection".
- Wrongfully recognizing an unknown person as known, and assigning it an identity. This is called a "false acceptance".

Correctly accepting a known person, and correctly determining his or her identity is referred to as "correct classification". From these values, we can infer a set of performance measures: The correct classification rate (CCR), the false classification rate (FCR) and the false rejection rate (FRR), which are the ratios of correct classifications N_{cc}, false classifications N_{fc} and false rejections N_{fr} with respect to the number of known persons N_{known} in the test set. Finally, the false acceptance rate (FAR), which is the ratio of false acceptances N_{fa} with respect to the number of unknown persons $N_{unknown}$ in the test set (see Eqs. 3.1, 3.2, 3.3, 3.4).

$$CCR = \frac{N_{cc}}{N_{known}} \tag{3.1}$$

$$FCR = \frac{N_{fc}}{N_{known}} \tag{3.2}$$

$$FRR = \frac{N_{fr}}{N_{known}} \tag{3.3}$$

$$FAR = \frac{N_{fa}}{N_{unknown}} \tag{3.4}$$

It follows that the sum of the correct classification, false classification and false rejection rates is equal to one.

$$CCR = 1 - (FCR + FRR) \tag{3.5}$$

These measures are commonly used in the biometrics literature for identification or verification tasks. As a difference to standard terminology, the term "impostor" will not be used here, because of its negative connotation. Indeed,

we do not consider occupants which do not belong to the set of known persons as trying to "deceive" the smart perception system. Rather, they are e.g. guests which have not previously been seen in the smart space, or do not visit it often enough for their identities to be permanently stored. Therefore, in the following, only the term "unknown person" will be used. In the remainder of this thesis, specifically for the evaluations made in Chapter 6, the task to be tackled is that of open set identification, which is why more detailed explanations concerning the problems posed for this task are given here.

There are a number of ways open set classification can be performed:

- Using a closed set classification algorithm with an additional class for unknown persons. This "unknown" class is trained with a large number of samples and generalizes the person class itself. The idea is that when presented with a corresponding test sample, person-specific classes will output higher scores than the general class, which therefore acts as an automatic threshold to detect and reject unknown persons.
- Performing verification for multiple classes. In a verification task, a separate classifier is trained for each person to distinguish only that person from all other (known and unknown) persons. This can be done by thresholding the output of a one-class classifier, by training a two class classifier (using samples from the target person for the positive class and all other samples for the negative class), and so forth. Open set identification is then made for a test sample by outputting the highest scoring identity for which verification was successful. The test sample is rejected as unknown if it could not pass the verification test for any of the known persons.
- Performing closed set classification with additional thresholding of the classifier outputs. In this case, classification is performed as in the closed set task, with the difference that the winner class is only accepted if its classification score is greater than a specified threshold. Otherwise, the test sample is rejected as unknown.

The main problem when using methods based on an "unknown" class is that often, sufficient data for training of that class is not available. One could remedy this in the multiple class verification case by using the data from all remaining known persons as negative samples for a given person or, in the case of a multi-class classifier with separate unknown class, by using the data from all known persons for training that class. While this may be practicable for parametric methods, such as Gaussian mixture model classifiers, it is not suitable for non-parametric methods, such as nearest neighbor or Support Vector Machine (SVM) classifiers. The reason is that these do not approximate a "mean" or "extension" for the unknown class, but rather more or less learn to classify each negative (in this case "known") training sample. As a consequence, test samples

for known persons score just as high for the unknown class as for person-specific classes, and no thresholding is longer possible.

The main difficulty of the threshold-based techniques is determining the threshold itself. Indeed, if it is chosen too high, all persons, including known ones, are eventually rejected. If it is chosen too low, eventually all unknown persons will be falsely accepted. Usually, no threshold will allow to retain all known test samples while rejecting all negative ones, and a compromise has to be found. In many cases, the threshold is chosen such that the false rejection and false acceptance rates are equal. The then achieved false rejection rate (which is equal to the false acceptance rate) is called the equal error rate (EER). Although the equal error rate can give an indication of the overall performance of a classifier, for many scenarios it may not be its optimal operating point. This can, for example, be the case when the number of unknown persons in the test set greatly outweighs the number of known persons. In this case, it may be desirable to choose a higher threshold to avoid large numbers of false acceptances, to the cost of rejecting a few more known persons.

As a consequence, open set classification performance, just as verification performance, is usually evaluated using Receiver Operating Characteristic (ROC) curves. These diagrams plot the correct and false classification ratios against that of false acceptances, as the classification threshold is shifted. Figure 3.6 shows an example ROC plot. In this plot, the CCR and FCR are shown as cumulative curves. This means that for a given FAR, the values of the CCR, FCR and FRR are successively added to create the respective curves, such that $CCR+FCR+FRR = 1$. The EER is shown as a diagonal line intersecting the CCR and FCR. The higher the intersection point, the better the equal error rate. The CCR value obtained for maximum false acceptance rate (usually at $FAR = 1$) is the accuracy obtained for closed set classification (even though this is not a "closed set" task, as unknown subjects are present in the test set). Note that the false rejection rate is not plotted, as it can be easily derived from the other curves (see Eq. 3.5). It is shown as the area above the FCR.

The following sections describe the algorithms used for face ID and speaker ID feature extraction, as well as the methods by which confidence measures are derived, that will later serve for thresholding in the open set identification task.

3.3.2 Features Extracted from Face Identification

The face recognition algorithm is based on the work of Ekenel et al. [49; 48]. It is an appearance-based technique using local Discrete Cosine Transform (DCT) features. It has been shown to provide accurate results under a variety of challenging conditions, including uneven lighting, shadows, partial occlusion [35]

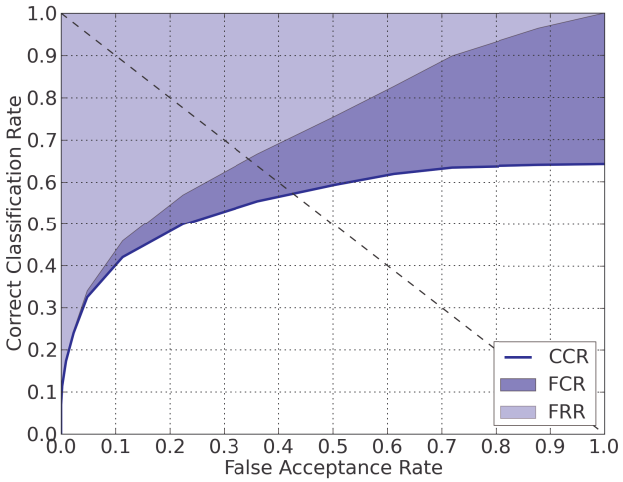


Figure 3.6: An example ROC curve for an open set identification case. The light blue area represents the false rejection rate.

and, perhaps most important in our case, low face resolutions, which is a common, almost unavoidable problem in an uncontrolled scenario [102; 103]. In the CLEAR seminar database, which will be used in the evaluations in Chapter 6, face sizes are as small as 15×15 pixels. Face recognition algorithms usually operate on much higher resolutions.

As for the majority of face recognition algorithms, the technique is designed to operate on frontal or near frontal views, mainly because of the difficulties in detecting and in aligning non-frontal faces. Although alignment of the face, for example by determining the eye regions, can greatly improve recognition performance, this is not done here as finding facial features at such low resolutions is not feasible. Instead, the detection window itself is taken as cropped face image, without further alignment. The recognition technique is based on local DCT features, which are extracted in the following way: First, the face area is resized to 32×40 pixels. This cropped image is then divided into a non-overlapping grid of 8×8 blocks, for a total of 20 blocks. For each block, the Discrete Cosine Transform is applied and the resulting DCT coefficients are ordered according to a zig-zag scan pattern. The first coefficient is dropped, as it only expresses the overall illumination level. From the remaining coefficients, the first 10 are selected to form the local feature vector for the block. Local feature vectors are further normalized to unit norm and concatenated to generate the global feature vector for the detected face (see Fig. 3.7).

Many recent approaches have shown the advantages of using local features, compared to holistic approaches, as they are more robust to lighting changes, occlusion, etc. A variety of techniques has been proposed that perform generic partitioning of the face region and DCT feature extraction [11; 67; 90]. The

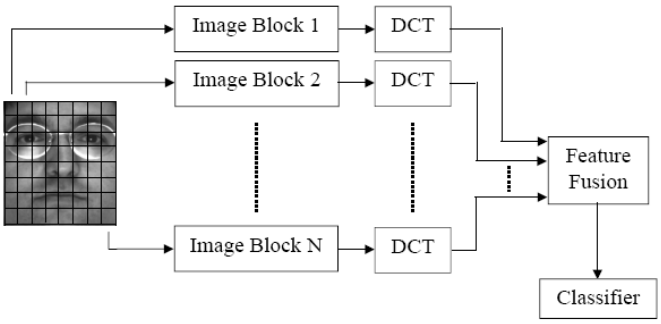


Figure 3.7: Local appearance-based face recognition. The face image is divided into 8×8 pixel blocks, DCT coefficients are computed for each block, normalized, and concatenated to a feature vector (Image taken from [37]).

details of the feature extraction are therefore not the main point of interest here and the interested reader is referred to [49; 48]. What is of concern to us at this point are the methods used for classification and for estimation of confidence measures in the identification, as these affect the way open set identification is later performed.

In this work, a non-parametric technique is employed. Feature vectors for a detected face are passed to a nearest neighbor classifier which is trained on sample vectors extracted using the same method as described above. The training vectors are gathered on a separate enrollment set for each of the known individuals, using labeled faces in multiple camera views. The feature vectors for all views are accumulated and used jointly for classification. The classification itself is performed by computing the distance of the test vector to each of the training vectors using the $L1$ norm. The set of resulting distances is then used to infer confidence scores. In the original implementation proposed by the authors for the CLEAR 2007 evaluations [36], this is done by sorting the distances in ascending order and min-max normalizing, such that the closest training vector has a score of 1 and the furthest a score of 0. The scores are further normalized to unit sum and the distance x between the best two scores used to compute a confidence weight:

$$\omega(x; \lambda) = 1 - \exp^{-\lambda x}$$

The authors term this weight the *distance-to-second-closest* metric. It should be noted that on various occasions, the difference between the classifier scores for the best two candidates (or the best n candidates) has been shown to be a valid means of estimating confidences. This is done e.g. by the same authors also in [36] for acoustic speaker identification using a-posteriori probabilities from a Gaussian mixture model classifier, or by Grosse et al. in [43] for temporal and multimodal fusion using n-best lists.

In contrast to the method proposed in [36], the confidence estimation here is done using the k nearest neighbors, as follows: The set of distances from the test vector to the training samples is sorted in ascending order and only the k nearest samples are retained. Their distances are min-max normalized, such that the first (closest) sample has a score of 1 and the kth sample a score of 0. These scores are now used as weights in a voting scheme to determine the identity of the test person. Each sample i votes for its corresponding identity $ident(i)$ with a voting weight ν_i proportional to its score:

$$\nu_i = \frac{1}{k} s_i, i = 1,\ 2 \dots k$$

with s_i the min-max normalized score for sample i and $ident(\cdot)$ the function that maps training samples to identities. The result is an n-best list of candidate identities $I = \{id_1, id_2 \dots id_n\}$ with $n \leq k$. The cumulative normalized scores for these identities are then computed as:

$$S_j = \sum_{i=1 \dots k, ident(i)=j} \nu_i,\ j = 1, 2 \dots n$$

with

$$\sum_{j=1 \dots n} S_j = 1$$

The identity with the highest score is output as the result of identification. Further, instead of using the distance between the best two scores, the value of the best score is simply used as confidence measure. This is viable since the scores sum up to 1. A value greater than 0.5, e.g., means that the score for the highest ranking identity outweighs that of all others. In the following, we will simply refer to the value of the cumulative normalized score (CNS) for the highest ranking identity as the "confidence" for the identification.

The reason for limiting the the min-max normalization and confidence estimation to the k nearest neighbors is that the number of neighbors used has a direct impact on the open set performance of the classifier. If k is too large, too many irrelevant samples are included, which decreases the quality of confidence scores. If k is too small, there are not enough candidate samples to estimate confidences with in the first place. This will be shown in detail in the evaluations section in Chapter 6.

As mentioned in Section 3.3, an alternative way to realize open set identification would be, e.g., to perform verification for multiple identities, or to specifically discriminate against unknown persons by training an "unknown" class. This

is however not possible for the CLEAR seminar database. Although the seminars figure both known and unknown persons, and an enrollment set exists for the known persons, there is no enrollment set of "unseen" persons that could be used for training of the unknown class. This is one of the reasons why a threshold-based method is adopted here. The second reason is that the confidence measures will be helpful in modality fusion, and also to some extent in temporal fusion, as will be shown in Chapter 6.

3.3.3 Features Extracted from Voice Identification

The voice recognition algorithm is based on the work of Jin et al. [54]. It is based on Mel Frequency Cepstral Coefficients (MFCC) as acoustic features and Gaussian Mixture Model (GMM) classification. The challenge in an unobtrusive identity tracking scenario is that recognition is to be made using solely distantly placed microphones. No close-talking microphones or lapel microphones that need to be worn by the users, but only omnidirectional microphones placed, e.g., on a table or microphone arrays mounted on the walls of the smart space are to be employed. As with face identification, the difficulties associated with far-field acoustic identification are multiple. They lie in segmenting speech, separating multiple speakers, and dealing with low signal to noise ratios, reverberations, laughter, etc. An additional problem, similar to unobservable faces which are turned away from cameras, is that identification can only be made while a person speaks, and also usually only for one person at a time. Crosstalk, which occurs when speech intervals from several persons overlap, is generally detrimental for speech segmentation, speaker localization and speaker identification.

The recognition algorithm itself is not designed for automatic speech segmentation or diarization (the process of association speech intervals to different speakers). It expects pre-segmented speech coming from exactly one speaker and performs recognition on whole speech segments. Inclusion of silence, crosstalk, or wrongful concatenation of utterances from alternating speakers into one segment for recognition decreases performance dramatically. This means that speech has to be detected and segmented properly beforehand by a different means, and that speech segments should be kept short, if possible, to reduce the chance of wrongful concatenation (e.g. because of insufficient pause). There are several ways to achieve this, including thresholding the audio signal in the power spectrum, performing explicit diarization [13; 88], filtering in the frequency domain to remove non-speech sounds [17], etc. The specific segmentation method used in this work will be described in the experimental setup section in Chapter 6. It delivers short audio segments of one or two seconds duration which are passed on for recognition.

The voice recognition algorithm works as follows: First, MFCC features are extracted from the audio signal. These are further normalized using reverberation

compensation and feature warping. This is necessary as distant-speech signals are corrupted by reverberation and background noise. The algorithm uses a modified version of the Cepstral Mean Subtraction (CMS) algorithm, designed to cope with the lengthy impulse response of the reverberation. Feature warping is done using a standard CDF matching technique (CDF stands for "Cumulative Distribution Function"). This transformation warps the distribution of the cepstral feature stream over a given time interval to a standard distribution and is useful, for ex., for reducing the non-linear effects occurring when using different channels. The result is a 13-dimensional warped MFCC feature vector, with reverberation compensation applied.

The feature vector is passed to a Gaussian mixture model classifier with a fixed number of mixtures (16 or 32) per speaker. A separate GMM is trained for each speaker on pre-segmented, labeled speech samples from an enrollment set using the expectation-maximization (EM) algorithm. For a given test segment, the log likelihood score for each GMM is computed as the sum of log likelihoods of all samples in the segment, given the model. The algorithm outputs the identity corresponding to the GMM with the maximum log likelihood score.

Again, the exact details of the recognition algorithm are not the main point of concern here and the interested reader is referred to [36]. Using MFCC features or variations thereof and GMM or SVM classifiers is common practice in the field of speaker identification[101; 20; 68; 17; 69], such that the proposed algorithm can be seen as a good representative of the current state-of-the-art. The interesting question is how the confidence measure for the made identification is derived in our case. Similar to the visual case, the scores of the n highest ranking identities are used. The difference here, is that parametric models (GMMs) are employed, such that distances to training samples need not be computed. Instead, the GMM log likelihood scores are directly used as scores for an identity. The k highest scores are first min-max normalized, and subsequently normalized again to unit sum. Finally, the maximum normalized score value is used as confidence measure for the identification. As for face identification, sufficient data for training of unknown classes is not available and the open set identification is realized by thresholding the confidence scores of the closed set classifier.

3.4 Spatial Localization and Composition of High-Level Features

The previous sections have described how identities and confidences are computed for faces detected in camera views and speech segments extracted from microphone signals. The identities and confidences are the main components of the high-level "face ID" and "speaker ID" features introduced in Section 3.1.

This section now shows how the feature descriptions are extended with information about the 3D scene locations of the so "detected" identities. The reason localization is made in 3D-space is to provide a common ground for later spatio-temporal association to tracks and feature fusion. This is in contrast to estimating, e.g., azimuths and distances relative to microphone arrays for audio features, and 2D image coordinates and sizes for visual features.

First of all, speaker and face ID features are tagged with an observation timestamp t_{obs}. For face ID features, this is the time of frame capture by the detecting camera. For speaker ID features, it is the capture time of the last sample in the considered speech segment. The time tag is necessary as the video and audio modalities are captured at wholly different rates (usually 15 to 30 frames per second for the visual case, 44.1 KHz for the audio case) and ID features are only observed at relatively few, irregular points in time. The synchronization of frame-level support maps, localized detections, face ID and speaker ID features is later made using these timestamps.

For face ID features, a color description for the identified person's upper torso is also extracted. This is done by defining a rectangular area directly under the face detection bounding box as a "virtual" upper torso detection (see Fig. 3.8). The area inside the so defined bounding box can then be used just as for the case of regular detections to compute the torso's mean color. The assumption made here is that the area under a detected frontal or near frontal face will almost always represent the region of the torso. This assumption may not hold, e.g., when the person whose face was detected is partially occluded by other persons, by laptops or other objects on the table he or she is sitting at, etc. Still, if the majority of views captured are uncluttered, the occasional outliers produced in these cases should be dealt with later in the color model building stage.

Finally, the localization information, for both for visual and acoustic cues, is calculated. It is composed of a 3D location $\vec{x} = (x, y, z)$ in a global reference frame, as well as a covariance matrix Σ_{loc} expressing the uncertainty in the found location. As observations in both cases are inherently noisy, the location can usually not be estimated exactly. This will be explained in detail in the following. In the later tracking stage, the covariance matrix will be used to perform data association in a probabilistic way and to reject unreliable location information.

3.4.1 Visual Localization

In the visual case, 3D localization is achieved by exploiting the expected real scene dimensions of the detected body region and camera calibration information. By using the detection bounding box center and width in a camera image, the line of view and distance from the camera to the object can be estimated

Figure 3.8: The estimated torso area relative to a detected face. Detected and identified faces represent very reliable cues for the presence of a person. They can be used to infer the position of the body and extract color or texture features in a somewhat more supervised manner.

and a 3D scene location computed. The explanation is first given for upper torso detections:

The calibration information for each camera is given relative to a global coordinate frame in the smart space. For each detection, the line of view from the camera to the detected torso in the scene can be estimated as the line passing through both the camera center and the detection center in the image plane. Similarly, the distance from the camera is estimated using the image width of the detection bounding box and an assumed real upper torso width (Considering the typical size of bounding boxes output by the detector, a scene width of 80cm was used here). From the direction and distance, a 3D point in the global coordinate frame is obtained. As detected upper bodies are not always tightly bounded, and since their orientation relative to the camera changes their apparent width in the image, this estimated 3D location comes with a certain amount of uncertainty, which is modeled in the covariance matrix Σ_{loc}. The eigenvalues of Σ_{loc} are the variances σ_x^2, σ_y^2 perpendicular to the line of view, and σ_z^2 along the line of view (see Eq. 3.6).

$$\Sigma_{loc} = R^T diag(\sigma_x^2, \sigma_y^2, \sigma_z^2) R \tag{3.6}$$

In Eq. 3.6, the rotation matrix R expresses the rotation from the global coordinate frame to a detection-specific coordinate frame with its origin at the camera

(a) **(b)**

Figure 3.9: The uncertainty in the 3D location of a detected person. Since the distance from the observing camera is estimated from the size of the detection bounding box, estimation errors are much larger along the line of view of the camera.

center, its z-axis being the line of view to the detected object and its y-axis parallel to the y-axis in the camera frame. R can easily be computed from the camera extrinsics and the offset of the detection center to the camera projection center. The exact computations involve standard methods in computer vision and projective geometry and will not be elaborated further here. Further details can be found in [46].

Figure 3.9 shows an example detection, its estimated 3D location and the localization uncertainty.

In the general case, the error in estimated distance is much higher than that in estimated direction from the camera. This is because small pixel differences in the detected bounding box can translate to large differences in estimated distance when the object is not near the camera. Further, the assumed actual width of the upper torso is, of course, an approximation which does not hold for all persons under all circumstances. This is even more true as the torso detector also fires for near frontal views, such that upper bodies may often be viewed under an angle. The apparent width in the image is then smaller, which translates to a larger estimated distance from the camera. For all these reasons, the location uncertainty is much greater along the line of view to the detected object than in directions perpendicular to the line of view. This is reflected in Σ_{loc} by a large value of σ_z compared to σ_x and σ_y. For the experiments in Chapter 6, the variances in Σ_{loc} are assumed fixed for upper torso detections and are determined empirically based on actual data.

The localization for detected faces is done using the same procedure and assuming an average scene width of 18 cm. For top view features, the distance from the camera is not estimated from evidence in the image. Rather, the tracked person center is assumed at a fixed height of 1m above the ground, such that

the line of view from the camera need only be intersected with the plane parallel to the ground plane at 1m height to obtain the 3D point location $\vec{x}$. The uncertainty covariance matrix Σ_{loc} is again computed using the same methods as described above.

3.4.2 Acoustic Localization

The method for localization of acoustic features is likewise more complex. It involves analyzing the signals from several microphone arrays distributed on the edges of the smart space. As an audio source can in principle be perceived by all microphones in the space, it is feasible to fuse the information at feature level and obtain a robust global estimate. There are several ways to accomplish this, as demonstrated for example in the CLEAR evaluation workshops [102; 103]. What makes these evaluations especially relevant for our case, is that they represent the first extensive evaluations of source localization techniques using distributed microphone networks (DMNs). In contrast to single microphone arrays setups (with linear arrays, spherical, etc), distributed networks present a greater challenge to localization techniques, as the far-field assumption does not hold between pairs of microphone arrays. Nevertheless the use of a distributed microphone network in a smart environment offers several advantages as it can cover a much larger space (it is not limited, e.g., to the area surrounding a meeting table) and allows to triangulate exact 3D positions (as opposed to relative directions or azimuths). State-of-the-art approaches utilize variations of the Generalized Cross Correlation function (GCC) to estimate time delays of arrival (TDoAs) of a signal between array microphones. These are either used in a filtering framework to infer the direction to the sound source [41], or the sound source location is directly derived from all observations by computing a Global Coherence Field (GCF) [96]. The algorithm used here is the one proposed by Gehrig et al. in [41]. It implements a Joint Probabilistic Data Association Filter (JPDAF) framework, to fuse the observations from several microphone arrays. The advantage of the JPDAF is that it is capable of maintaining several sound source tracks to which observations are associated in a probabilistic way. This, in principle, allows the filter to switch more quickly between alternating sound sources, a case which commonly occurs e.g. in conversations. It also makes it more robust to occasional sources of noise, as these are not associated to and do not disturb the track of the main target.

The algorithm works as follows: First the TDoAs and corresponding correlation values are calculated for all possible microphone pairs within each array by calculating the GCC Phase Transform (GCC-Phat) [60; 87] of the frequencies below 8KHz. The estimation is made 25 times per second and the corresponding correlation values are thresholded using a predetermined threshold. The JPDAF is then fed with one measurement vector for each time instant and

microphone array which is made up of the TDOAs of those microphone pairs of that array with correlation values above the threshold. The measurement vector is only used for position estimation if it has at least 2 elements. The algorithm maintains an iterated extended Kalman Filter (IEKF) for each internally tracked target and evaluates the joint probabilities for associating events to tracks. The possibly of an observation originating from a target is given by the target's innovation covariance matrix. The state update for each target is then be made separately using the PDAF update rule [16]. The selection of the active sound source out of the maintained targets is done by choosing the target with the smallest error covariance matrix volume.

The result of source localization is the 3D scene location of the current most active sound source, as well as an associated uncertainty covariance matrix. If the uncertainty in the estimation is too large, the result is immediately regarded as invalid and discarded. Likewise, if a sound source is estimated for a time interval where no speech identification could be made, the result is discarded. This is to avoid tracking noise sources such as dragged feet, printers, etc.

Localization is performed at a much higher rate than identification, as the former is made every 40ms and the latter is the result of analysis on time intervals of several seconds. In a final step, the location of the identified speaker is therefore taken as the average of all successful localizations within the identification window, if the variance of estimated locations in this time frame is below a specified threshold. Otherwise, no location information is given. One should note this can lead to cases where a speaker identification is possible, but localization is not. In those cases, the speaker ID cue can later not be associated to a person track based on spatial correlation, but can well serve as a hint that the concerned person is actually present in the smart space. In the next chapter, we will see how the integrated tracking approach makes use of this information.

3.4.3 High Level Feature Description

With the information obtained from 3D localization, the composition of high-level features can now be completed. The description for detection features, as well as for top view features is composed of a 3D location $\vec{x}$ with associated localization uncertainty Σ_{loc} in the global coordinate frame, as well as the mean color description of the detected or approximated torso region $col = (\mu_R, \mu_G, \mu_B)$.

$$det = (t_{obs}, \vec{x}, \Sigma_{loc}, col)$$

$$top = (t_{obs}, \vec{x}, \Sigma_{loc}, col)$$

Face ID features carry a large amount of information, being composed of a 3D location $\vec{x}$ with uncertainty Σ_{loc}, an estimated upper body color col, the index

of the highest ranking identity id_V and an associated identification confidence $conf_V$.

$$fid = (t_{obs}, \vec{x}, \Sigma_{loc}, col, id_V, conf_V)$$

Speaker ID features are composed of the index of the recognized identity id_A, an identification confidence $conf_A$ and, if available, the estimated origin in the scene of the identified speech.

$$sid = (t_{obs}, \vec{x}, \Sigma_{loc}, id_A, conf_A)$$

3.5 Summary

This chapter presented the various low-level and high-level features used in the identity tracking framework. They include less reliable or descriptive but frequently observable features used mainly to keep track of identified persons, as well as highly detailed and specific, though harder to observe features used for accurate detection, color model initialization and identification. The features are extracted using state-of-the-art algorithms and techniques adapted to smart environment scenarios involving multiple users. The extraction is done using only distantly placed cameras and microphones in an unobtrusive, opportunistic way, without requiring the attention or cooperation of users.

The extracted features are:

- Foreground feature maps in the form of integral images,
  $fg(x, y)$
- Color feature maps in the form of R, G and B-channel integral images,
  $col(x, y) = (col_R(x, y), col_G(x, y), col_B(x, y))$
- Localized detection features,
  $det = (t_{obs}, \vec{x}, \Sigma_{loc}, col)$
- Localized blob tracking features from top view cameras,
  $top = (t_{obs}, \vec{x}, \Sigma_{loc}, col)$
- Localized face ID features,
  $fid = (t_{obs}, \vec{x}, \Sigma_{loc}, col, id_V, conf_V)$
- Localized and non-localized speaker ID features,
  $sid = (t_{obs}, \vec{x}, \Sigma_{loc}, id_A, conf_A)$

4 Probabilistic Multiple Identity Tracking using Irregularly Observable Features

The main idea guiding the design of the identity tracking framework is to opportunistically integrate reliable but sparsely available cues for identification whenever they become available, and to keep tracking recognized persons in the absence of such. Audio-visual tracking and identification features of varying accuracy, specificity, frequency and level of abstraction are made available by the feature extraction steps described in Chapter 3. These features arrive in an asynchronous and often very irregular way. This raises the need for a flexible fusion technique that integrates highly varied, partially incomplete and possibly very sparse information.

The goal is to track and identify multiple persons evolving freely in a smart space. The identification is to be made for the open set case, which means that not all persons in the space are known beforehand. Let $\mathcal{I} = \{id_1, id_2 \dots id_n\}$ be the set of known identities. These are the identities of persons for which voice or face models have been trained in a priori, for example because they frequently visit the smart space, belong to a privileged group or are in some form or another the focus of interest (for example the main speaker in a meeting or the presenter in a seminar). Let $\mathcal{P}$ be the set of persons evolving in the smart space at a given point in time. Let $\mathcal{P}_\mathcal{F}$ be the subset of persons in $\mathcal{P}$ whose identities are known. They will subsequently be referred to as "focus persons" or simply "known persons". Let $\mathcal{P}_\mathcal{U}$ be the set of remaining "unknown" persons. The task is then to:

1. Detect and track all persons in the smart space. This includes automatically estimating the number of persons, determining their exact locations and keeping consistent trajectories throughout the duration of their attendance.
2. Distinguish known focus persons from unknown persons.
3. Identify all focus persons.

A person's state can be seen as comprised of his or her location, color of clothing, identity, pose (e.g. standing or sitting), body orientation, focus of attention, and

so forth. For our purposes, the location and identity are of main relevance, with the color of clothing used as a means to an end. They are considered as hidden variables to be jointly inferred in a probabilistic estimation process. The state of a person $P \in \mathcal{P}$ is therefore defined as

$$S_P = (\vec{x}, COL, ID).$$

The location variable $\vec{x}$ is eventually subject to frequent and fast change, as the person moves around, and needs to be continuously updated. The color of clothing COL is not expected to change (at least not in the here considered time interval), though the appearance of colors may change slightly in time, e.g. due to illumination changes, or may differ depending on the observing sensor. It needs to be learned using the sequence of observations associated to the person. The identity of a person also does not change, although errors in estimation of the identity are to be expected. This is why it also needs to be estimated using a sequence of observations. Further, while it is possible for several persons to wear the same color of clothing, identities are unique and cannot be shared, such that identity variables ID need to be estimated jointly for all persons.

The performance of the identity tracker will be measured based on two criteria: One is the accuracy in tracking all persons. The other is the accuracy in determining the presence, the location and the identity of focus persons. Depending on the type of application envisioned, the two criteria may not be of equal importance. An example case is that of a smart meeting room which only needs to perceive the actions of the meeting organizers (to keep protocol, automatically display presentation slides, etc.) while ignoring the members of the audience. In this case, the former criterion may be of small interest. Similar requirements are conceivable for distributed perceptual environments in office buildings or shopping malls, with large amounts of day-by-day visitors. In such application scenarios, it may be infeasible or unnecessary to track the movement of all persons. This is one of the reasons the identity tracking framework is designed specifically to cope with partial or faulty tracking information. The measures for person tracking and "identity tracking" performance will be discussed in detail in Chapter 5.

In the following, the proposed algorithm for joint estimation of multiple identities, the Joint Identity Tracking (JIT) filter, is described. It recognizes the set of known persons in the smart environment, while considering their locations as additional information which may or may not be available. The person locations are inferred in a hybrid Bayesian filtering framework, combining sequential Monte-Carlo (particle) approximations with grid-based techniques. The determination of identities present in the smart space and their association to tracked locations are made by probabilistic filtering in a specially designed joint identity space. As detection and tracking features as well as inferred information about identities are considered potentially flawed, all available cues are used to estimate the number of persons in the scene.

4.1 Localization

The goal of localization is to detect and track all persons in the smart environment. The location $\vec{x}$ of a person is estimated with respect to a global coordinate frame and represents the (x, y, z) coordinates of the centroid of the person's head. The reason for using the head centroid as reference point is that it is much less ambiguous than, e.g., the body center, which may be hard to determine depending on the body pose, or the ground location of the feet, which may be unobservable in many cases. It also makes it easier to evaluate tracking performance, as the ground truth for evaluation, the 3D head location, can be easily obtained by annotating the heads in multiple camera views and triangulating. Although the tracking is made in three dimensions, the evaluation of the tracker will only be made using the (x, y) coordinates on the ground plane, with information about person height considered of secondary importance. This will be explained in more detail in Chapter 6.

As inputs to the localization, all features described in Chapter 3 are utilized. These include foreground and color support maps from one or more cameras, torso detections, top-view blob tracks, detected and identified faces, as well as identified and localized speech. The support maps and detections can be considered "classical" tracking features and are more or less regularly observable as long as the tracked persons are in a camera's field of view. Localized face and speaker ID features, although less frequently observable, are also very useful for tracking as they represent highly reliable types of detections. Although as in many visual or audio-visual tracking approaches, using a greater number of observing cameras increases the tracking accuracy, the algorithm is designed to function with variable numbers of cameras, with its accuracy degrading gracefully as the number of cameras decreases. In the extreme, it "tracks" the locations of alternating speakers based only on the acoustic modality, using no camera information at all.

For the probabilistic tracking of persons using multiple camera views, a certain class of algorithms, particle filters [52], has been shown to be particularly successful [112]. Particle filters are Monte Carlo approximations of the Bayesian filtering framework. Due to their non-parametric nature, they are easily applicable to filtering problems with non-linear, non-Gaussian state dynamics, are robust to noise or outliers, and can be flexibly used with varying types of sensors and features [85; 21; 63; 24]. For these reasons, particle filtering has been chosen as the base of the localization framework, with several extensions proposed to increase robustness with respect to very low observability of features. Low observability can manifest itself on several aspects. It is, for example, inherent in the audio modality as persons take turns speaking, such that features cannot be extracted for all of them continuously. It is given for the visual modality when faces that are turned away from cameras or persons regions that are occluded or

presented in non-standard poses cannot be detected. It is given when coverage of the smart space is not complete and persons leave the field of view of sensors, or even when single modalities are wholly unavailable. The proposed extensions are in the form of a special class of "unobserved" particles, a modification of the sampling proposal distribution, and the incorporation of grid-based approximation techniques. They will be described later in this section, following a description of the implemented particle filter and of associated observation models.

4.1.1 Particle Filter Framework

In Bayesian filtering, the state of a dynamic system is estimated from a sequence of noisy observations. In the case of person tracking, the system state can be, e.g., the location of a person, his or her orientation, velocity, height, etc., and the observations are the features extracted by various sensors. The state of a person at time t is represented by a random variable x_t and the filter estimates the probability distribution over x_t in a recursive process. Let $\{z_1, z_2 \dots z_t\}$ be the set of observations up to time t. The filter estimates the posterior density of x_t given all observations

$$p(x_t|z_{1:t})$$

Computations are made under the Markov assumption: The probability of the current state given the previous state is conditionally independent of earlier states

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

Similarly, the observation at time t is conditionally independent of all other states given the current state.

$$p(z_t|x_{t:1}) = p(z_t|x_t)$$

Using these assumptions, the probability distribution for x_t is recursively estimated as:

$$p(x_t|z_{t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|z_{t-1})\, dx_{t-1} \tag{4.1}$$

and

$$p(x_t|z_t) \propto p(z_t|x_t)p(x_t|z_{t-1}) \tag{4.2}$$

Equation 4.1 is referred to as the belief propagation or prediction step. Equation 4.2 is the filtering or update step [52].

In particle filtering, the probability distribution $p(x_t|z_{1:t})$ is approximated by a set of weighted samples (the particles), $\left\{x_t^{(n)}, w_t^{(n)}\right\}_{n=1}^{N}$, with $\sum_{n=1}^{N} w_t^n = 1$. Each particle represents a sample of the random variable x_t, i.e. a hypothesized state. The weights are called importance factors and determine the importance

of each sample. The sequential Bayes filter updates can be realized in a variety of ways. The most common is through Sampling Importance Resampling (SIR) [52]. Let $\left\{x_{t-1}^{(n)}, w_{t-1}^{(n)}\right\}_{n=1}^{N}$ be the particles and weights representing the posterior distribution at time $t-1$. The particle set representing the posterior at time step t is then generated as follows:

- Resampling: particles are sampled N times with replacement from the set of particles $x_{t-1}^{(n)}$ according to the weights $w_{t-1}^{(n)}$. This produces a set $\left\{x_{t-1}^{(n)'}\right\}_{n=1}^{N}$.
- Propagation according to system dynamics: $p(x_t|x_{t-1} = x_{t-1}^{n'})$ is used to sample each new particle x_t^n. This corresponds to the prediction step of Bayesian filtering.
- Observation scoring: Each new particle x_t^n is assigned a weight proportional to the observation likelihood $p(z_t|x_t = x_t^n)$. This corresponds to the update step.

In the resampling step, new particles are drawn according to a so-called proposal distribution. In importance resampling, this distribution is given by $p(x_t|x_{t-1})p(x_{t-1}|z_{t-1})$. That is, the posterior from time step $t-1$, after propagation, is used as prior for the update step at time t. In Section 4.1.6, a modification of this proposal distribution will be introduced, to account for cases in which the prior at time t cannot explain the observation z_t. Before this, the implemented observation models $p(z_t|x_t)$ and dynamic model $p(x_t|x_{t-1})$ are first explained in Sections 4.1.2 and 4.1.3.

In the following tracking framework, the set of person states is referred to as $\mathcal{S} = \{S_1, S_2, \ldots, S_{N_P}\}$. As mentioned above, the state vector of one person is comprised of his or her location, upper torso clothing color and identity

$$S_i = (\vec{x_i}, COL_i, ID_i),\ i = 1, 2 \ldots N_P$$

For the purpose of tracking, the location vector $\vec{x_i} = (x_i, y_i, z_i)$ is extended with velocity information (dx_i, dy_i, dz_i)

$$S_i = (x_i, y_i, z_i, dx_i, dy_i, dz_i, COL_i, ID_i),\ i = 1, 2 \ldots N_P$$

The use of the velocity information will be explained in Section 4.1.3. The first six parameters of the state vector represent the dynamic part of the person state and are continuously updated in the particle filtering process. The last two parameters are static variables which are incrementally estimated from the sequence of observations $\{z_1, z_2, \ldots, z_t\}$.

In the multiple person tracking case, the particle filter framework offers the possibility of performing the estimation for all person in a joint state space

$$S = (x_1, y_1, z_1, dx_1, dy_1, dz_1, COL_1, ID_1, x_2, y_2, \ldots, dz_{N_P}, COL_{N_P}, ID_{N_P}).$$

The disadvantage in using joint state spaces, however, is that as the dimensionality of the state space increases with the number of persons, the number of required particles grows exponentially. To avoid this, special techniques for managing the increased dimensionality are required. These techniques can greatly increase the complexity of the tracker, which is why a joint state space approach was not used here. Instead, person locations and clothing colors are estimated using individual state spaces, which amounts to using a separate particle filter per person. The advantage is that computations are straightforward and fast. The disadvantage is that track exclusion has to be achieved by separate means, to avoid several tracks converging on the same target. This will be described in Section 4.1.4. Person identities, on the other hand, are estimated in a joint state space described in Section 4.2.

4.1.2 Observation Models for Scoring

The filtering algorithm maintains a separate track (in fact, a separate particle filter) for each person in the smart space. Let $\mathcal{T} = \{T_1, T_2, \ldots, T_{N_T}\}$ be the set of tracks. Again, the state vector of a track is defined as $S_i = (x_i, y_i, z_i, dx_i, dy_i, dz_i, COL_i, ID_i)$. In the update step of the filter, the particles for each track are scored (their weights are updated) based on the likelihood of the current observation given the track state $p(z_t|S_{i,t})$. The observation vector z_t is itself composed of the foreground and color support maps, detected upper torsos, top view detections and face ID features from the set of observing cameras, and the speaker ID features for time t. Let

$$\{cam^m\},\ m = 1 \ldots M$$

be the set of observing cameras,

$$fg_t^m(ix, iy)$$

$$col_t^m(ix, iy)$$

the foreground and color support maps for cam^m (Here ix and iy are used to distinguish image coordinates from world coordinates x, y and z.),

$$det_t^{m,nd},\ nd = 1 \ldots Ndet_t^m$$

$$top_t^{m,nt},\ nt = 1 \ldots Ntop_t^m$$

$$fid_t^{m,nf},\ nf = 1 \ldots Nfid_t^m$$

the detection, top view and face ID features for cam^m and

$$sid_t$$

the speaker ID feature at time t, if available.

Then the observation likelihood is defined as the product of the component likelihoods of individual feature types (in the following, the subscript t is omitted for simplicity)

$$p(z|S_i) = p_{fg} \cdot p_{col} \cdot p_{det} \cdot p_{top} \cdot p_{fid} \cdot p_{sid} \cdot p_{sb}$$

with

$$p_{fg} = \prod_{m=1}^{M} p_{fg}^{m} = \prod_{m=1}^{M} p(fg^{m}(ix, iy)|S_i)$$

$$p_{col} = \prod_{m=1}^{M} p_{col}^{m} = \prod_{m=1}^{M} p(col^{m}(ix, iy)|S_i)$$

$$p_{det} = \prod_{m=1}^{M} p_{det}^{m} = \prod_{m=1}^{M} \prod_{nd=1}^{Ndet^{m}} p(det^{m,nd}|S_i)$$

$$p_{top} = \prod_{m=1}^{M} p_{top}^{m} = \prod_{m=1}^{M} \prod_{nt=1}^{Ntop^{m}} p(top^{m,nt}|S_i)$$

$$p_{fid} = \prod_{m=1}^{M} p_{fid}^{m} = \prod_{m=1}^{M} \prod_{nf=1}^{Nfid^{m}} p(fid^{m,nf}|S_i)$$

$$p_{sid} = p(sid|S_i)$$

and p_{sb} expressing the likelihood given the boundaries of the smart space itself. This component will be explained in detail further below.

Many particle filter-based tracking approaches in the literature use the sum of component likelihoods (scores) instead of the product. This helps avoiding floating point errors due to the multiplication of small decimal numbers. It also allows to define "weights" for the scores of individual components, which are used in the computation of a weighted average. The weights then serve to adjust the importance of specific feature types and are set manually or automatically [86]. Here, the product of components is taken, with "importance weights", if any, realized by a careful definition of the individual conditional probabilities. An advantage of the product rule is the implementation of penalties for states inconsistent with (part of) the observation vector, simply by setting one of the component likelihoods to a small value (in the extreme, to zero). This would be the equivalent of using "negative" scores in the case of the sum rule, which can be problematic as the total score may then as well become negative.

Although separate state spaces are employed, the scoring or update step is not performed independently for each track. This is due to the fact that occlusions caused by one track can influence the observation likelihood of another track. Also, the association of localized features to tracks, which directly influences likelihoods, is made probabilistically based on all current tracks states. The occlusion model will be discussed in detail in Section 4.1.5.

The observation models used to derive the observation likelihoods for the different feature types will now be explained in detail.

"Smart Space Boundary" Model

Since the tracking of identities is to be made only inside the smart observation space, it makes sense to penalize particles that leave its physical boundaries. This is done here simply by using the bounding cuboid SB of the space. The upper and lower boundaries of SB are set to sensible values representing the meaningful area in which person heads are to be found (head locations, which are approximated by the particles, should e.g. not be below 90cm or above 2m). Apart from eliminating gross errors, this increases computation speed by reducing the size of the state space. Let $S^n = (\vec{x}^n, \vec{dx}^n, COL^n, ID^n)$ be the state vector for the nth particle. Particle scores are then updated using the likelihood function

$$p_{sb}(z_t|S_t^n) = \begin{cases} 1 & \vec{x}^n \in \text{SB} \\ 0 & else \end{cases}$$

Strictly speaking, the so defined function does not model an *observation likelihood*, as it is not dependent on z_t. Describing it as part of the observation model is a generalization: One can conceive of much more complex, possibly dynamic boundary functions, that would also consider the placement of movable objects, such as tables, desks, etc., forming absolute exclusion zones for person. The location of these objects may itself be part of the observation vector, such that the boundary model is again dependent on z_t.

In the current form, boundary exclusion could be seen as part of the prediction step: The prediction for the particle mass is corrected by setting the weights of particles falling outside the smart space boundaries to zero. Since the computation is performed directly following the prediction step, this is equivalent to using a somewhat more elaborate dynamic model.

$$p(S_t|z_t) \propto p(z_t|S_t) \int p(S_t|S_{t-1}) p(S_{t-1}|z_{t-1})\, dS_{t-1}$$

$$= \prod_{feat} p_{feat}(z_t|S_t) p_{sb}(z_t|S_t) \int p(S_t|S_{t-1}) p(S_{t-1}|z_{t-1})\, dS_{t-1}$$

$$= \prod_{feat} p_{feat}(z_t|S_t) \int \left[p_{sb}(S_t) p(S_t|S_{t-1})\right] p(S_{t-1}|z_{t-1})\, dS_{t-1}$$

with $p_{feat}(z_t|S_t)$ the observation models for actually observed features.

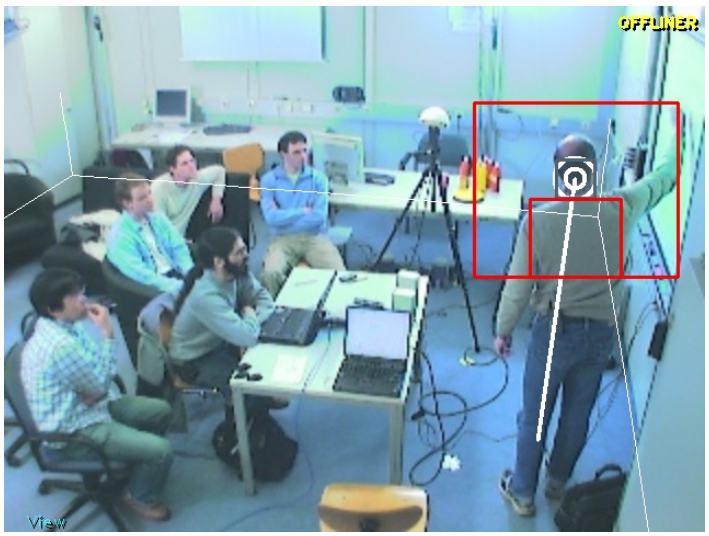


Figure 4.1: The inner and outer bounding boxes used to estimate the foreground likelihood score for a particle. The highest score is achieved if there is full support in the inner box, and no support in the immediately surrounding region.

Foreground Observation Model

The foreground observation likelihood is computed as the difference of ratios of foreground pixels inside the person area defined by the particle state, and of foreground pixels in the immediately surrounding area. To compute $p^m_{fg}(z|S^n)$, a bounding box B_{in} is first constructed by projecting the estimated location and size of the person into the corresponding camera view. A second bounding box B_{out} is then constructed, centered around B_{in} and with 5 times its area (see Fig. 4.1.

The foreground ratio R_{in} is computed by dividing the number of foreground pixels inside B_{in} by the bounding box area. Likewise, the foreground ratio for B_{out}, excluding the area occupied by B_{in}, is computed. This is efficiently done using the integral image representation of the foreground support. Next, the ratio difference between the two areas is computed as $R_{diff} = R_{in} - R_{out}$. Its value ranges from -1 to 1. Assuming perfect foreground segmentation, a value of $R_{diff} = 1$ is achieved when the foreground support in the target region is exactly constrained to the estimated person region. A value of -1 represents the worst case: that any location surrounding the estimated region is better suited to support the track hypothesis. Intermediate values are obtained if the person region is not well bounded or when other persons are in the immediate surrounding, making the localization less reliable. Now, let $0 < c_{eq} < 1$ be a non-negative value representing the (non-zero) likelihood that a person is present in

the target region in the case $R_{in} = R_{out}$. The observation likelihood is then defined as the piecewise linear function

$$p(fg^m(ix, iy)|S^n) = \begin{cases} c_{eq} + R_{diff}(1 - c_{eq}) & R_{diff} \geq 0 \\ (1 + R_{diff})c_{eq} & else \end{cases}$$

The parameter c_{eq} can be used to weight the importance of the foreground feature. If c_{eq} is set very high, the lack of adequate foreground support is not strongly penalized and particle scores are mostly influenced by other features. When it is set very low, the foreground feature is very restrictive. In any case, the best and worst case scenarios will result in values of $p(fg^m(ix, iy)|S^n)$ of 1 and 0, respectively.

Color Observation Model

The observation likelihood for the color feature $p^m_{col}(z|S^n)$ is computed for each particle only if a color model is available for the corresponding view. The color model for the mth view, COL^m consists of a mean μ^m_{col} and diagonal covariance matrix Σ^m_{col} and is only available if it has been learned in in previous observations of the track. More details on initializing color models are given in Section 4.1.9. Similar to the foreground feature, the observation likelihood is computed here using the projected bounding box B_{in} of the upper torso into the corresponding image. Using the integral image representation for each channel, the mean color μ_B is quickly computed for the area inside B_{in} and used to derive a color score

$$s_B = \mathcal{N}(\mu_B, \mu^m_{col}, \Sigma^m_{col})$$

with $\mathcal{N}(\cdot, \mu^m_{col}, \Sigma^m_{col})$ the multivariate Gaussian function centered at μ^m_{col} and with covariance matrix Σ^m_{col}. The observation likelihood is then obtained as

$$p(col^m(ix, iy)|S^n) = c_{eq} + s_B(1 - c_{eq})$$

If no color model is available for a specific view, the corresponding observation likelihood is set to $p^m_{col}(z|S^n) = c_{eq}$. The reason for enforcing a minimum value c_{eq} for the color likelihood is that since color models are initialized and adapted in an unsupervised way, the influence of a failed color match on the overall observation likelihood should be limited. Here again, c_{eq} can be seen as a parameter adjusting the importance of the color similarity feature.

Detection and Top View Observation Models

Localized upper body detections are scored based on spatial proximity and color similarity. The spatial component $Sp_{corr}(n)$ for the detection feature $det_{m,nd} = (t_{obs}, \vec{x}, \Sigma_{loc}, col)$ extracted from cam^m and the particle with location $\vec{x}^n$ is given by

$$Sp_{corr}(n) = \mathcal{N}(\vec{x}^n, \vec{x}, \Sigma_{loc})$$

The color similarity component Col_{corr} is computed using the track's color model for the corresponding view $COL^m = (\mu^m_{col}, \Sigma^m_{col})$, if available

$$Col_{corr} = \mathcal{N}(col, \mu^m_{col}, \Sigma^m_{col})$$

The resulting observation likelihood is then

$$p(det^{m,nd}|S^n) = c_{obs} + (1 - c_{obs})Sp_{corr}(n)Col_{corr}$$

with c_{obs} a minimum likelihood value for particles that do not correlate (spatially or in color space) with the detection feature. In this way, a smooth association of localized observations to tracks is implemented: For each detection, the particle scores of tracks that do not coincide with the detection are set to the minimum non-zero value. Particles that correlate well receive a higher weight. To simplify the likelihood computation, a gating function is used that immediately sets the weights of particles with a distance greater than 3σ from the detection center to c_{obs}.

For top view features, the observation likelihoods $p(det^{m,nd}|S^n)$ are computed in the same manner.

Face ID Observation Model

Localized face ID features are scored based on spatial proximity and color similarity for the associated upper torso region, just as for detection features. The correlation of the observed identity with the modeled identity, however, is not used in the scoring stage. The reason for this is that single-frame face ID estimates and single-segment speaker ID estimates are expected to be quite noisy, especially in the presence of unknown persons. Rather than penalizing particles based on the identity correlation, effectively modifying the a-posteriori distribution of track locations, single identity estimates are associated to tracks, accumulated, and a probability distribution for *identity locations* is inferred. This will be explained in detail in Sections 4.2 and 4.3.

In summary, the observation likelihood for face ID features is defined as

$$p(fid^{m,nf}|S^n) = c_{obs} + (1 - c_{obs})Sp_{corr}(n)Col_{corr}$$

with $Sp_{corr}(n)$ and Col_{corr} defined as above, and the parameter c_{obs} used for the same reasons.

Speaker ID Observation Model

Similarly, the observation likelihood relating to localized speaker ID features is computed based on spatial proximity.

$$p(sid|S^n) = c_{obs} + (1 - c_{obs})Sp_{corr}(n)$$

Non-localized speaker ID features are not considered in scoring.

4.1.3 Prediction and Resampling

The resampling of particles is made independently for each track using standard Sequential Importance Resampling (SIR). A new set of particles is drawn from the particle set $\{x^n\}_{n=1}^N$ according to the importance weights given by the observation likelihoods $w^n = p(z|S^n)$. For this, particle weights are first normalized such that $\sum_{n=1}^N w^n = 1$. After the resampling step, a new set of particles $\{x'^n\}_{n=1}^N$ is obtained, with equal weights $w'^n = 1/N$.

The following propagation (or prediction) step updates the new particle states (in essence their location) using a simple linear dynamic model: The locations are deterministically updated using the particle velocities

$$\vec{x}_{t+1}^n = \vec{x'}_t^n + \vec{dx}_{t+1}^n$$

while the velocities (initially set to 0) are probabilistically updated using Gaussian noise

$$\vec{dx}_{t+1}^n \propto \mathcal{N}(\vec{dx'}_t^n, \Sigma_{dx})$$

with $\Sigma_{dx} = diag(\sigma_{dx}, \sigma_{dy}, \sigma_{dz})$ a diagonal covariance matrix. The variances $\sigma_{dx,dy,dz}$ are parameters for the prediction step and are set to model the reasonable range of velocity change in an indoor environment. In this way, the dynamics of the state space are effectively constrained to 3 dimensions, dx, dy and dz, as the location (x, y, z) is deterministically inferred.

4.1.4 Mutual Track Exclusion

Since the prediction and filtering steps described so far are performed independently for each track, we must take care that several tracks do not converge on the same target. This problem is well known in the field of multi-target tracking and several solutions have been proposed [70]. One solution involves using estimated track centers [23]. In such an approach, the most likely location for each track is computed, e.g. as the weighted average of its particle locations, and particles from neighboring tracks are penalized when they lie inside a certain region around the track center. The problem with such approaches is they assume the

underlying probability distribution about the person location to be unimodal, while in the particle filter approach, a non-parametric, possibly multimodal distribution is modeled. If, e.g., the belief about a person's location presents two distinct peaks at opposite ends of the room, the computed track center lies in the middle of the room and may wrongfully perturb another track. Another track exclusion method involves the use of Markov Random Fields [58; 62]. These are employed in an iterative process during the prediction step to gradually approximate the conditional distribution $p(S_t^1, \ldots, S_t^{N_T} | S_{t-1}^1, \ldots, S_{t-1}^{N_T})$. The computations can be relatively complex, as they involve single particles, but offer the advantage of preserving the non-parametric assumption.

Here, a much simpler method is used, based on grids. Non-parametric Bayesian filtering using grid-based methods has already been used with success in a number of applications. The advantage of grid-based methods is that they are simple to implement and can represent arbitrary distributions over a discretized state space. The downside is that their computational and space complexity grows exponentially with the number of dimensions. Here, the complexity problem is circumvented by applying only 2-dimensional grids to solve specific tasks, such as mutual exclusion on the ground plane, while the main filtering steps are accomplished in the particle filter framework.

The method works as follows: First, the ground plane of the smart space is subdivided into a regular, non-overlapping grid of cells $C_{i,j}$. The width w_{cell} of a cell (which is equal to its height) is a parameter to the algorithm (throughout this work, a width of 10cm is used, as it provides a fine enough tessellation for person tracking). Three types of occupancy maps are computed, by accumulating in each cell the normalized weights of all particles falling inside the cell:

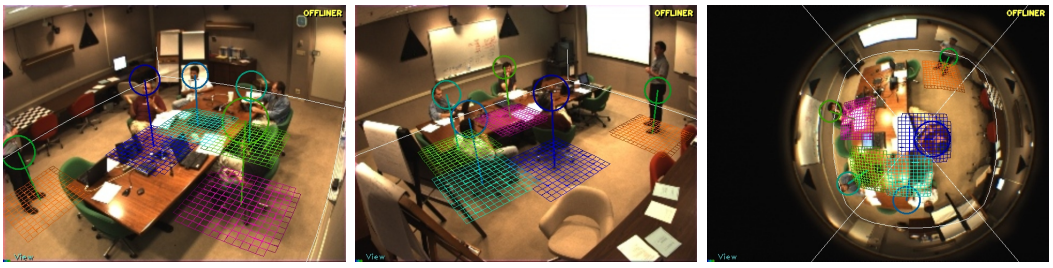
- An individual occupancy map for each track T_k, $ocp_k(i,j)$ with

$$ocp_k(i,j) = \sum_{n=1\ldots N, x^{k,n} \in C_{i,j}^k} w^{k,n}$$

- The total occupancy map for all tracks, $ocp_{tot}(i,j) = \sum_{k=1}^{K} ocp_k(i,j)$.
- From the former two, the complementary occupancy map for each track, $ocp_{!k}(i,j) = ocp_{tot}(i,j) - ocp_k(i,j)$

Cells with an accumulated weight greater than zero are referred to as *active cells*. Only they need to be considered in the track exclusion process. The computed occupancy maps are used as a means to realize a coarse and fast clustering of the particle mass, and will be useful also for generation of the tracker output or for implementation of a fast occlusion mechanism.

Mutual track exclusion is realized as follows: The main constraint to be realized is that a specific region on the ground plane can only be occupied by one person at a time. As a consequence, the sum of accumulated normalized weights for



(a) **(b)** **(c)**

Figure 4.2: The occupancy maps calculated upon the grid-based discretization of the smart space. Cells are color coded according to the different tracks. The higher the occupancy of a cell, the brighter its intensity.

one track inside a certain region around a cell can serve as a penalty term for other tracks. The exclusion region RC for a cell $C_{i,j}$ is approximated as the square with center (i, j) and with a radius R_{excl} expressed in numbers of cells . Region-based occupancy maps $Ocp_k(i, j)$ are then computed with

$$Ocp_k(i,j) = \sum_{C_{x,y} \in RC} ocp_k(x,y)$$

This can be done efficiently again by applying the integral image technique on the previously defined grids. Likewise, the complementary occupancy maps $Ocp_{!k}(i, j)$ are computed. Figure 4.2 shows example occupancy maps for multiple tracks as seen from various camera views.

In the resampling step of the particle filter, a penalty term $excl^{k,n}$ is defined for each particle of a track T_k by using $Ocp_{!k}(i, j)$: Let $C_{i,j}$ be the cell occupied by particle $x^{k,n}$. Then

$$excl^{k,n} = 1 - min(1, Ocp_{!k}(i,j))$$

That is, as the particle mass from other tracks in the region around $C_{i,j}$ increases, the penalty term tends toward 0. Before resampling, the exclusion penalty is multiplied with the particles' normalized weights, followed by a renormalization to unit sum. The effect is that single particles from different tracks tend to repel each other, which prevents the accumulation of their probability mass on a local scale.

4.1.5 Occlusion Handling

Several approaches in the visual multiple target tracking literature have shown that specifically modeling occlusions between targets can greatly improve tracking performance [70; 62]. This is why an occlusion modeling technique for visual

observations is also implemented here. It computes a 2.5-dimensional, camera-relative representation of target-related occlusions in the scene, by using the above introduced occupancy maps. For each camera cam^m, a fast, dynamic programming algorithm derives an occlusion factor $occ^m(i,j)$ for each active cell of the previously described grid. These occlusion factors are then used to modify the observation likelihoods for features extracted from the corresponding camera view.

Computing probabilistic track occlusion for discretized scene locations is of course a simplification. Another possibility is to calculate image-based occlusion maps, such as proposed e.g. in [62]. The advantage is that such maps are much more detailed as the occlusion can be estimated at the pixel level. The disadvantage is their computational complexity. The advantage of the method proposed here is that it can be computed very fast and used effectively with all types of features extracted from the same camera view. A major difference to the method in [62] is that only one occlusion map is generated for all tracks per camera, whereas in [62], individual maps are computed for each track. The method works as follows:

For each camera cam^m, the occlusion factor for an active grid cell $C_{i,j}$ (active meaning that $ocp_{tot}(i,j)$ is not zero) is computed using the total occupancy of cells lying between it and the camera. Let $\{C_1, C_2, \ldots, C_N\}$ be the set of cells lying on the line of view lov from $C_{i,j}$ to camera cam^m and let $C_0 = C_{i,j}$. Let $0 < ocl_n^m < 1$ be a factor expressing the amount of occlusion caused by cell C_n. Then, ocl_n^m is calculated as the sum of the occupancies of C_n and its neighboring cells NC_n. NC_n is defined as the set of cells lying on the line perpendicular to lov passing through C_n, with distance to C_n smaller than a specified occlusion radius R_occl. The reason is that due to the size of persons being larger than the size of a cell, persons hypotheses lying in the neighborhood of C_n also contribute to occluding the line of view (see Fig. 4.3).

The occlusion factor $occ^m(i,j) = occ^m(0)$ for cell $C_{i,j}$ can then be recursively computed using

$$occ^m(n) = ocl^m(n) + occ^m(n+1)$$

$occ^m(i,j)$ is truncated to values between 0 and 1 and the recursion is broken off as soon as the accumulated value reaches 1 (once the line of sight is fully obstructed by one person, it is irrelevant how many more persons still stand behind him or her). The occlusion factors for all cells are calculated using an efficient dynamic programming algorithm. The values computed for cells on a line of view are stored, such that they need not be recomputed for another line of view that passes through them. This makes for an extremely fast recursive computation.

The algorithm described up to this point calculates occlusions using a 2D grid and neglecting completely the height of persons in the scene. When cameras

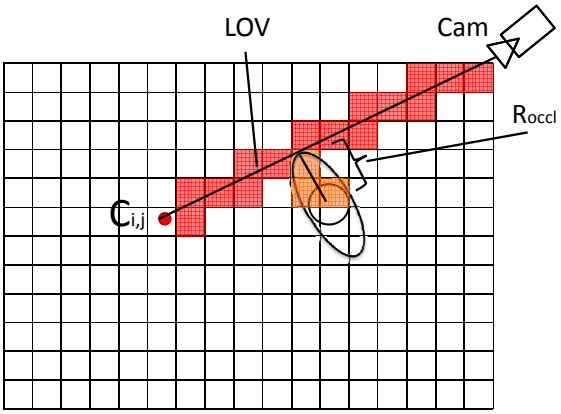


Figure 4.3: The computation of occlusion factors based on grid discretizations. The occupancies of cells lying on the line of view to the camera are accumulated in a recursive fashion. The line of view is considered blocked completely if the cumulative occupancy exceeds a value of 1.

are placed at a sufficient height, however, it is well possible for persons to stand behind each other, relative to one camera, without occluding each other. The algorithm is therefore extended to include approximate height information in the following way: For computation of the occlusion factor $occ^m(i,j)$, a visibility cone V is first projected from the camera center to $C_{i,j}$, such as to cover the vertical range of areas potentially occupied by upper torsos at $C_{i,j}$ (see Fig. 4.4). For this, upper and lower bounds, b_{high} and b_{low}, for the upper body area are defined (Here, b_{high} and b_{low} are set to 2m and 1m, respectively). The reason generic bounds are used is that occlusion is calculated for a whole cell, and not for each particle hypothesis (with associated height) inside the cell. This is, of course, an approximation, which however proves to be sufficiently precise for our purposes. When performing the recursive computation, V is used to estimate a correction term $f^m_{corr}(n)$ for the occupancy of each cell C_n on the line of view to the camera. $f^m_{corr}(n)$ is defined as the percentage of the conic section inside C_n that lies between b_{high} and b_{low} (see Fig. 4.4). The idea is that if the visibility cone is not intersected, the view on the relevant parts of $C_{i,j}$ is not occluded and $f_{corr} = 0$. Using the correction term, the recursive formula is redefined as

$$occ^m(n) = (f^m_{corr}(n) ocl^m(n)) + occ^m(n+1)$$

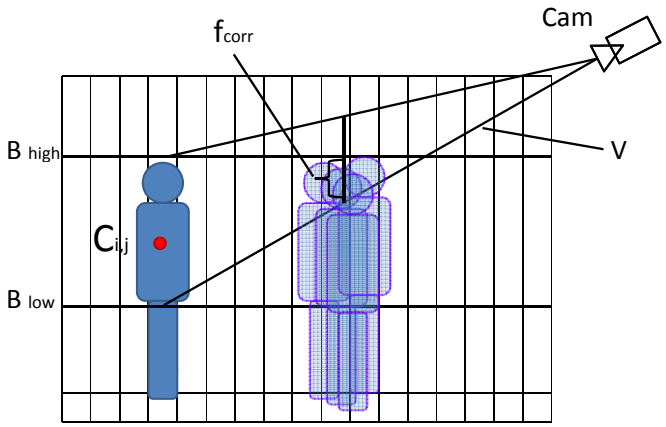


Figure 4.4: The extension of the grid-based occlusion model to include height information. The contribution of cells on the line of sight to the camera is weighted according to the portion of the visibility cone they are expected to maximally intersect.

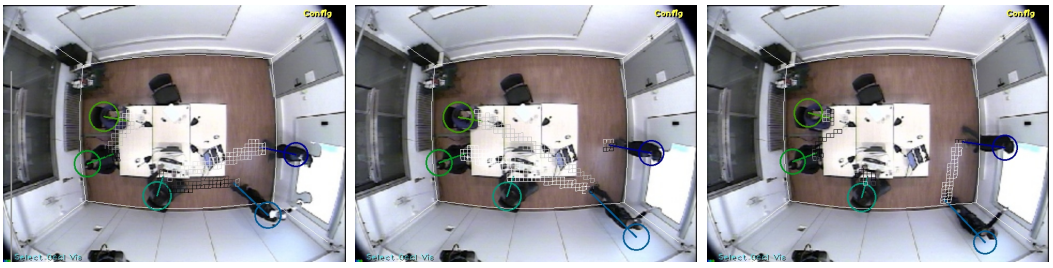
Figure 4.5 shows example occlusion maps generated for several observing cameras in a multiple track scenario.

The algorithm is computationally very efficient for two reasons:

- Occlusion is only computed for active cells. Since the number of active cells is smaller than or equal to the number of particles, usually only a fraction of the total grid is considered.
- The dynamic programming algorithm avoids computing the occlusion factor for the same cell twice. When a cell on the line of sight to the camera for which the occlusion factor was already computed is reached, the recursion returns.

Using the occlusion map for camera cam^m, the observation likelihoods for track T_k are modified as follows: Let $C_{i,j}$ be the cell occupied by particle x^n. Let further $p(feat_M^m(ix, iy)|S^n)$ with $feat_M \in \{fg, col\}$ and $p(feat_L^{m,nl}|S^n)$ with $feat_L \in \{det, top, fid\}$ be the observation likelihoods for feature maps and localized features extracted from camera cam_m, respectively. The modified observation likelihood for feature maps is then

$$p'(feat_M^m(ix, iy)|S^n) = occ^m(i, j)c_{obs} + (1 - occ^m(i, j))p(feat_M^m(ix, iy)|S^n)$$



(a) **(b)** **(c)**

Figure 4.5: The grid-based occlusion maps for three different corner cameras, as seen from a same top view point. Lighter cells represent areas of strong occlusion, while darker ones represent less occluded ones. Occlusion factors are calculated only for the cells that are actually occupied by particles in a fast recursive algorithm. The direction of the occlusion shadows in the separate images indicates the origin of the observing camera.

and the likelihood for localized features becomes (see Section 4.1.2)

$$p'(feat_L^{m,nl}|S^n) = c_{obs} + (1 - c_{obs})Sp_{corr}(n)Col_{corr}(1 - occ^m(i,j))$$

That is, in both cases particles that are completely occluded receive a likelihood score of c_{obs}. The parameter c_{obs}, which has already been introduced in Section 4.1.2, can be seen as a base likelihood for particles which cannot be scored as they are too distant to be associated to localized observations, occluded by other particles, and so forth. It will be explained in more detail in the next sections.

4.1.6 The Issue of Observability

First, the concept of "unobserved" particles and tracks is introduced. An unobserved particle represents a hypothesis about the target state which cannot be verified by observation because it cannot be detected acoustically and is either inside the visual occlusion zone of another target, out of view of all sensors, or outside of the smart space altogether. An unobserved track is a track for which all particles are unobserved. This can happen, e.g., when camera coverage is low and a person clearly passes from a covered to an uncovered area. The motivation behind the concept is as follows: When tracking multiple targets, observations may only be available for one of them at a time (e.g. when tracking based only on speech). No observation can be made for other tracks during this time. Yet, we wish to keep track of their positions until new observations are available. This is common practice when tracking single targets, e.g. using Kalman filters, or when tracking multiple targets, in conjunction with data

association techniques such as the Joint Probabilistic Data Association Filter (JPDAF), graph-based Bayesian belief nets, etc. The problem when using common particle filter implementations is that they rely on regular state updates for each target. If there are large gaps in observations of a target, the particle mass for its track will spread uncontrollably, due to the noise introduced in the probabilistic resampling step, which may quickly render the prediction given by system dynamics useless. This is especially true when tracking targets with highly unpredictable dynamics, such as humans, which may change their direction, stop moving and pause for lengthy periods of time, unexpectedly start to move again, walk back and forth, etc. For this reason, the assumption is made here that when a person cannot be observed, his or her state does not change. More specifically, we assume the location to be unchanged, with the uncertainty in this location growing as time passes. When a new observation can be associated to this person, the update is made based on the last observed location and velocity. This is common practice when using e.g. the aforementioned filtering techniques. The difference here is that this mechanism is implemented at the particle level: For unobserved particles only, the propagation step is skipped (their locations and velocities are not updated), while states for observed particles, on the other hand, are propagated as usual. When a localized observation is later made which is inconsistent with the current distribution of tracks, the weights and the previously known positions and velocities of their unobserved particles are used to update the belief in the tracks' states, effectively sampling new particles at the location of the observation. This will be explained in detail in Section 4.1.7.

By doing the data association and update at the particle level, the non-parametric nature of the underling target density is preserved and a hybrid system is obtained, where each sample of the particle filter acts as separate filter, which is updated probabilistically using Bayesian inference. As a simplification, we will speak in the following of the "unobserved particle mass" or "unobserved probability mass", meaning the sum of the unobserved particles' normalized weights.

A useful application of the concept of a partially unobserved particle mass is in determining the probability of the presence of a target inside the bounds of the smart space, i.e. the confidence in the existence of the track itself. Indeed, while the particles evolving inside the boundaries of the smart space jointly represent the hypothesis that the target is present, one may also compute a probability for the opposite hypothesis, that the target is outside of the space. A simple way to do this is proposed here: The hypothesis that a person is not present in the space is represented by an additional particle, the "exterior particle" x^{ext} with weight w^{ext}, which by definition is unobserved. It is handled similarly to other unobserved particles: since it has no valid location, it is neither propagated, nor resampled. In the filter update step, it is scored using a fixed value expressing

the likelihood of the observation, since the observation could not have been caused by the exterior particle. The scoring is done as follows:

- For support maps $feat_M \in \{fg, col\}$, the observation likelihood is

  $$p(feat_M^m(ix, iy)|S^{ext}) = c_{obs}$$

  and therefore

  $$p_{feat_M}(z|S^{ext}) = \sum_{m=1}^{M} c_{obs}$$

- For localized visual features $feat_L \in \{det, top, fid\}$, it is

  $$p(feat_L^{m,nl}|S^{ext}) = c_{obs}$$

  and

  $$p_{feat_L}(z|S^{ext}) = \sum_{m=1}^{M} \sum_{nl} c_{obs}$$

- For localized speaker ID features, it is

  $$p_{sid}(z|S^{ext}) = c_{obs}$$

In this way, the score of x^{ext} equals that of a completely occluded particle (see Section 4.1.5). The likelihood score of other particles can therefore only become smaller than that of x^{ext} if they are at least partly unoccluded and are penalized at some point by a foreground or color support map. After scoring, w^{ext} is normalized jointly with all other particle weights, such that

$$\sum_{n=1}^{N} w^n + w^{ext} = 1$$

The normalized weight of the exterior particle, the "exterior probability", can then be used as a track confidence, representing the overall validity of the track. If $w^{ext} > \sum_{n=1}^{N} w^n$, the visible evidence is not enough to support the existence of the track in the smart space, and it should be deleted. In the field of person or object tracking, it is common practice to use confidence thresholds, e.g. based on the average or the maximum observation likelihood for a track. Here, the thresholding is achieved in a more elegant and flexible way using the parameter c_{obs}.

4.1.7 Correcting the Prior on Uncovered Observations

Two types of features are used in the proposed tracking framework: support maps and localized features. Evaluating the observation likelihoods for support maps requires sampling the state space and evaluating the positions hypothesized by the samples. This is done using the usual update rules for particle filtering. Localized features, such as detections or preprocessed blob tracks, on the other hand, offer direct hints about the target distribution. They can be seen as temporary peaks in the observation likelihood which do not need to be indirectly found by propagating the particle mass, but can be scored directly based on e.g. proximity or other similarity criteria (see Section 4.1.2). Indeed, one of the problems inherent in particle filtering is that if the state space is not sufficiently well sampled or if peaks in the observation likelihood are too narrow, local maxima may be missed entirely. The use of localized features, i.e. regions of the state space which have been found to be of interest in a preprocessing step, allows to circumvent this problem (This is the case, e.g., for top view features, which are coherent peaks in the foreground support map of the top view camera, found through low-level preprocessing). They can be used to score the particles in their vicinity, but they can also draw the attention of the tracker to regions which should, but are currently not being sampled. Localized features are commonly used e.g. in Kalman filter-based trackers to update the belief about a target's state or to initialize new tracks for observations which do not match any of the current targets. Likewise, spatio-temporally localized observations are often used in conjunction with Dynamic Bayes Nets (DBNs) for multi-target tracking under occlusion (see e.g. [119]). Here, the particle filtering algorithm is extended to offer the advantages of both approaches.

For this, the notion of "coverage" of localized features is introduced. The coverage of a localized feature $feat_L \in \{det, top, col, fid, sid\}$ is computed as the sum of the normalized weights of all particles that can be associated to it, weighted by their spatial proximity

$$cov(feat_L^{m,nd}) = \sum_{n=1}^{N} Sp_{corr}(n) w^n$$

with $Sp_{corr}(n)$ the spatial proximity score as described in Section 4.1.2.

In the following, we will use the term "covered observations" to refer to localized features for which the coverage exceeds a certain threshold T_{cov}. Likewise, "uncovered observations" are those with a coverage lower than T_{cov}.

Uncovered observations can be caused by one of the following:

- Existing targets that have been temporarily lost or could not be observed for a certain time, and that got rediscovered. Also, targets that are badly tracked, and for which the actual observation constitutes a hint at their

possible true location. In this case, the particle mass of corresponding tracks is at least partially unobserved.

- New targets entering the smart space. For these targets, new tracks should be created.
- Errors, noise, faulty detections, etc. If possible these should be recognized as such and ignored.

The remainder of this section deals with the first of these three cases, while the others are handled in Section 4.1.8.

The unobserved particle mass of a track represents hypotheses for the track's state (its location) which could not be confirmed by observation for at least the previous time frame $t-1$. It also represents the probability for the hypothesis that the tracked target is actually not present in the smart space. In both cases, there is some uncertainty to the locations hypothesized by the unobserved particle mass. When an observation is made, which cannot be explained by the currently hypothesized states, the probability that it originated from unobserved particles, taking into account their location uncertainty, is computed. A new particle x^η is then sampled at the observed location, with weight w^η proportional to the total probability for the unobserved particle mass. Let x^n be an unobserved particle for track T_k with associated weight w^n. Let τ be the elapsed time since x^n was last observed and $\vec{x}^n_{t-\tau}$, $\vec{dx}^{\,n}_{t-\tau}$ its last observed position and velocity. Let also $\vec{x}^{obs}_t$ be the location of the uncovered observation and $\Delta\vec{x} = \vec{x}^{obs}_t - \vec{x}^n_{t-\tau}$. The probability of x^n being observed at $\vec{x}^{obs}_t$ is estimated by computing the acceleration $\vec{a}$ required for x^n to bridge the distance $\Delta\vec{x}$ in time τ, considering its previous velocity

$$\vec{a} = \frac{\Delta\vec{x} - \vec{dx}^{\,n}_{t-\tau}\tau}{1/2\tau^2}$$

Since the possible values of $\vec{a}$ are limited by the system dynamics expressed by $\Sigma_{dx} = diag(\sigma_{dx}, \sigma_{dy}, \sigma_{dz})$ (see Section 4.1.3), the sought probability is given by

$$match^n_{obs} = p(\vec{x}^n_t = \vec{x}^{obs}_t | \vec{x}^n_{t-\tau}, \vec{dx}^{\,n}_{t-\tau}) = \mathcal{N}(\vec{a}, \vec{0}, \Sigma_{dx})$$

For the "exterior particle" x^{ext} from Section 4.1.6, for which the location and velocity are undefined, it is simply taken as

$$match^{ext}_{obs} = 0.5$$

The total probability for the unobserved particle mass of track T_k is then obtained as

$$match_{obs} = \sum_{n=1}^{N} match^n_{obs} w^n + match^{ext}_{obs} w^{ext}$$

with N the number of unobserved particles. Let W_k be the total weight of unobserved particles for track T_k. For each uncovered observation $obs = feat_L^{m,nd}$, a new sample x^η is then inserted at location $\vec{x}^{obs}$, with weight $w^\eta = match_{obs,k}$, and the weights of all unobserved and newly created particles is renormalized such that their sum equals W_k. The coverage for the observation is also updated as

$$cov'(obs) = min(1, cov(obs) + \sum_{k=1}^{K} match_{obs})$$

In essence, the above described procedure modifies the belief in the tracked target's state prior to the filter update step, by shifting some of the track's unobserved particle mass toward uncovered observations. Applying modifications to the proposal distribution used in resampling, e.g. dependent on the current observation vector, is an active topic of research, and many variations have been proposed in the literature [78; 112]. Here, the idea is taken a bit further by the introduction of an "unobserved" and therefore "uncertain" probability mass which allows to modify also the belief $p(x_t|z_{t-1})$ prior to observation scoring.

4.1.8 Track Creation and Deletion

Observations that are still uncovered after the previously described matching stage, either because the unobserved mass of available tracks is insufficient or because the matching score is too low, are hints for the presence of a previously undetected person and trigger the creation of new tracks. On the other hand, tracks which are not supported by observation evidence for a certain amount of time are deleted. This section explains the mechanisms for track creation and deletion in detail.

New tracks are created on uncovered detection, top view, face ID or localized speaker ID features. As the feature extraction step is error-prone, a temporary "scout" track is first created, which still has to be validated by further evidence. The criteria for validating a scout track are as follows:

1. The exterior weight w^{ext} of the track, as described in Section 4.1.6, must smaller than the sum of other particle weights.
2. The average foreground likelihood per view $\sum_{n=1}^{N} p_{fg}^m(z|S^n)w^n$ must exceed a threshold Th_{fg} in 60% of available views.
3. The average color likelihood per view $\sum_{n=1}^{N} p_{col}^m(z|S^n)w^n$ must exceed a threshold Th_{col} in 60% of available views
4. The track's observability must be high enough throughout all views. Let $o^{m,n}$ be 1 if particle x^n is observable using features from cam^m, and 0 else.

Then the track observability is given by $\sum_{m=1}^{M}\sum_{n=1}^{N} o^{m,n}$ and must exceed a threshold Th_{obs}.

If the above criteria are met for a minimum amount of time τmin_{valid}, the scout track is validated and its position will subsequently be output by the tracker. If they are not met before a maximum time limit τmax_{valid}, the scout track is deleted.

Track deletion is accomplished using the same criteria, with the difference that criteria 2 and 3 must be valid only for 40% of available views. If the criteria cannot be met by a valid track, it is not immediately deleted, but rather temporarily invalidated and kept alive for a short period of time $\tau_{sustain}$. The idea is that tracks that become unreliable for short periods of time, e.g. due to temporary occlusion, should be allowed to reacquire their target. If the validation criteria can again be met within $\tau_{sustain}$, the track is again validated, otherwise it is deleted. Note that tracks for persons recognized as known (tracks for which the accumulated identification confidence exceeds a specified threshold) are kept alive for a significantly longer period of time τ_{known}. The reason is that they are much less likely to be spurious tracks caused by false detections, etc., and should not be dropped as quickly.

4.1.9 Unsupervised Color Model Learning

The color models for the set of tracks $\{T_k\}_{k=1}^{K}$ are updated during the scoring step of the filter. The color model for each track T_k consists of a set of view-dependent models $\{COL^m\}_{m=1}^{M}$ which are initialized and updated based on the color description *col* of localized features extracted in the corresponding camera views $\{cam^m\}_{m=1}^{M}$. These view-dependent models are defined as

$$COL^m = (\mu_{col}^m, \Sigma_{col}^m),\ m = 1, \ldots, M$$

with μ_{col}^m and Σ_{col}^m the mean and variance of mean colors accumulated for view m. As described in Chapter 3, the color described in the features and modeled in the tracks is the estimated mean color of a person's upper torso. Localized features are mapped to tracks in a probabilistic data association step, in which each feature is assigned to exactly one track for color adaptation. While localized features may contribute to the scoring of particles from multiple tracks, color learning is performed only for the track that best correlates with the observation. This is because a soft mapping of color observations to several tracks will quickly lead the degradation of color models. A hard decision, on the other hand, favors the creation of well defined, discriminative color models.

The association to tracks is made based on spatial coverage and color similarity. Let $\{x^n\}_{n=1}^{N}$ be the set of particles for track T_k with weights w^n and location vectors $\vec{x}^n$. Let also $feat_L^{m,nl}$ with $feat_L \in \{det, top, fid, sid\}$ be a localized

feature originating from cam^m, COL^m be the corresponding color model from T_k, if available, and $Sp_{corr}(n)$, Col_{corr} the spatial proximity and color similarity functions from Section 4.1.2. Then the correlation factor for $feat_L^{m,nl}$ and T_k is

$$corr(feat_L, k) = \sum_{n=1}^{N} Sp_{corr}(n) Col_{corr} w^n$$

If no color is associated to the observation (as is the case e.g. for speaker ID features), the value of Col_{corr} is set to 1 and the correlation is computed based only on spatial proximity. The association is then made to the track T_κ with the highest correlation value

$$\kappa = argmax_{k=1,\ldots,K}(corr(feat_L, k))$$

The reason why separate color models are used for each view is to avoid making strong assumptions about color constancy. As no inter-camera color normalization is performed, colors may have a substantially different appearance from one camera to the next. This is especially the case for top view cameras, e.g., where the ground plane forms the background in most of the image. The appearance of the upper body is also influenced by other factors, such as lighting, shadows, (the backs of chairs for sitting persons,) etc. This is why individual models are preferred.

When observations from a specific view are insufficient to build a dedicated model, an "average" color model COL^{avg} is used, with mean μ_{col}^{avg} and variance Σ_{col}^{avg} estimated using all available observations from other views. Though this model is less accurate than dedicated ones, it is sufficient for an initial, rough approximation. As soon as observations become available for the concerned view, a dedicated model is built and subsequently used.

The color models for views from wall-mounted cameras are updated only when upper torso or face detections are available. The reason is that the detections offer a reliable, though not always observable cue for estimating the boundaries of the torso (as compared to e.g. motion-based segmentation or constant adaptation at the likeliest target location). This avoids learning in background colors, which can quickly lead to persistent faulty color models. In comparison, color models for top views are updated at a much higher rate, as the top view features are in fact foreground blob tracks which, in the best case, are available at framerate. The learned models are less reliable, though, as the top view features are more error-prone and offer a much less precise segmentation of the torso region. This is another reason for the color-based feature filtering described in Chapter 3.

Empirical tests have shown, though, that a constant update of color models in all views using a very small learnrate can improve the quality of color models. This update is done by projecting the estimated, most likely upper torso location

into each view, and computing the mean color from the resulting bounding box. This is useful, e.g., for views in which no direct detection of the upper body can be made (e.g. because of an atypical pose), and where its appearance differs from the calculated mean of other views. If the track location can be maintained using other features and other views, the color model for this view is eventually acquired.

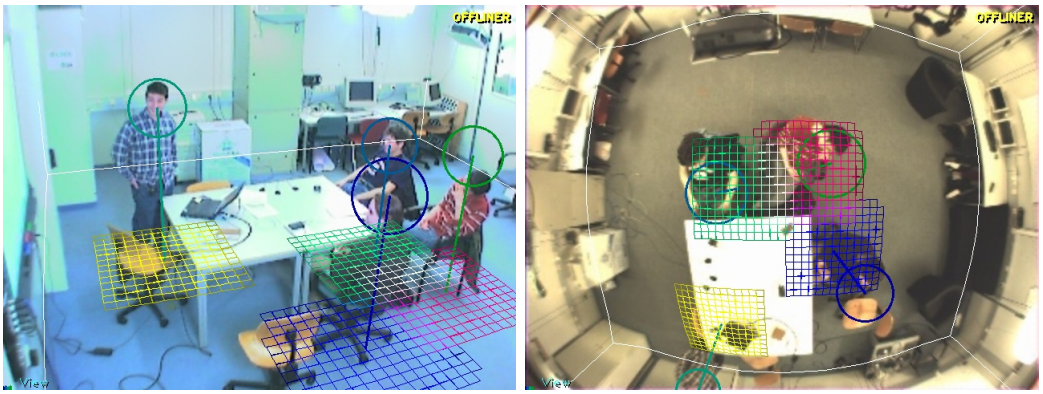
One should note that the unsupervised learning of color models, in the absence of a clear, reliable cue for deciding the relevance of observations, is a very challenging task [103]. Many sources of error exist:

- Faces can be detected, for which the associated upper torso is partly or completely occluded by objects or other persons.
- Inaccuracies in distance estimation for localized features can lead to faulty track associations.
- Inaccurate tracks can also lead to faulty associations in individual views, etc.

Though the solutions adopted here allow to cope with many of the mentioned problems, the automatic bootstrapping of classifiers or initialization of models for person tracking under realistic conditions is a tough problem and constitutes a still open topic of research.

4.1.10 Output Track Locations

The output of the particle filter localization process are the 3D locations of the tracked occupants' heads in the global coordinate frame. The location of a track T_k needs to be estimated from the locations of its particles. There are several ways to accomplish this, e.g. by using the weighted mean of particle locations, by using the location of the highest scoring particle, etc. The problem with the former method is that it breaks the non-parametric assumption about the target distribution. If the particle mass is split into two equally important clusters, more or less the mean of the cluster centers is output, which may not be a valid location at all. The problem when using the latter method is that the output location may be highly unstable, as it is based on single samples. Other viable alternatives involve the clustering of particles, which is what is done here. The method performs a fast clustering of particle locations, using the occupancy maps $Ocp_k(i,j)$ described in Section 4.1.4. The grid cell $C_{i,j}^{k,max}$ occupied by the maximum weight particle from track T_k is determined and a weighted average of particle locations is computed using only particles which lie in cells within the neighborhood of $C_{i,j}^{k,max}$. The neighborhood is taken as the exclusion region R_{excl} for cell $C_{i,j}^{k,max}$, as defined in Section 4.1.4. The method has shown to provide very stable position estimates while staying usable in the case of highly



(a) **(b)**

Figure 4.6: The output of the localization step. Again, track-specific occupancy grids are used to efficiently cluster particles and infer a locally smoothed average.

clustered output distributions. Figure 4.6 again shows an example of occupancy maps used to compute output track locations.

4.2 Identification

The previous sections have explained the filtering process by which the locations of occupants of the smart perceptual space are estimated. This includes initializing and deleting tracks, estimating track locations and confidences, and building person-specific color models. This section now describes how models for the identities of tracked persons are built from the speaker ID and face ID features gained throughout an observation sequence.

Just as for the case of color features, speaker ID and face ID features are matched to tracks in a probabilistic association step, and identity models are updated only for the track with the highest correlation value. Before the correlation function can be explained, we need to define the identity model of a track and the information extracted from each feature type.

The identity model ID, referred to in short as the "ID model", is a discrete, non-parametric probability density function (pdf) over the space of known identities. One can consider the values $ID(i)$ for discrete indices i as confidence values for the respective identities, with $\sum_{i=1}^{N_{ids}} ID(i) = 1$. Neighboring values in ID are independent of each other, as they represent concrete identities, such that no "mean" or "variance" for an identity can be inferred. If all values in the ID pdf are equal, the modeled identity is considered unknown. If a distinct peak in ID

exists, the identity is assumed known and the height of the peak can be seen as a measure of the confidence in the identification.

The identity-related information contained in the face and speaker ID features consists of the index id of the most likely identity and a confidence value $conf$. This information is first transformed to the same representation as for a track's ID model, namely to a discrete probability density function, before correlation is made. This is done by building a probability density function idf for the observation, such as

$$idf(i) = \begin{cases} conf & i = id \\ \frac{1-conf}{N_{ids}-1} & else \end{cases}$$

The result is a discrete pdf peaked at the index of the highest confident identity and with equal confidence for all other identities.

The motivation for computing idf in this way will now be explained further. In Sections 3.3.2 and 3.3.3, the output of the classification process is given as an n-best list of identities $\{id_1, \ldots, id_n\}$ with normalized confidence scores $\{conf_1, \ldots, conf_n\}$. The question one could ask is why this array of confidence scores is not directly used as discrete identity pdf (with confidences for remaining known identities set to 0). One of the reasons is that the classifiers used are nearest neighbor and GMM classifiers trained for the closed set case. When unknown persons are presented for classification, or when the decision for known persons is not clear, the classifiers typically output several identities with similar confidence. The n best confidences are then artificially brought to the range $[0, 1]$ by min-max normalization, and later renormalized. As a result, the values for the first n best scores are actually dependent on the size of the n-best list (or better said on the value n) and should not be considered as actual probability measures for the corresponding identities. When a definite peak in the output n-best list exists, on the other hand, the 2nd best, 3rd best identities, etc., are often random and are rather the results of discretization (for example due to limitations in the amount and type of training data). For these reasons, only the best scoring identity and its normalized score are further considered.

An additional reason, as will be described later in Section 4.2.1, is that the chosen method allows us to warp the confidence value of the identification before performing audio-visual fusion.

After the probability density function idf for an observation has been computed, the association to tracks is made based on spatial coverage and identity correlation. Let $\{x^n\}_{n=1}^N$ be the set of particles for track T_k with weights w^n and location vectors $\vec{x}^n$. Let also $feat_I \in \{fid, sid\}$ be a localized face or speaker ID feature and, ID^k be the ID model for T_k, if available. The spatial proximity component for single particles is again given by $Sp_{corr}(n)$, as defined in Section 4.1.2, and the identity correlation Id_{corr} is computed using the Bhattacharyya coefficient $B_C(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)}$ (the Bhattacharyya coefficient

is a measure of similarity between two discrete probability density functions. For details, the reader is referred to [55])

$$Id_{corr} = B_C(idf, ID^k)$$

The correlation factor for $feat_I$ and T_k is then

$$corr(feat_I, k) = \sum_{n=1}^{N} Sp_{corr}(n) Id_{corr} w^n$$

In the case of non-localized speaker ID features the correlation factor is computed as

$$corr(sid, k) = \begin{cases} Id_{corr} & Id_{corr} \geq \tau_{sid} \\ 0 & else \end{cases}$$

That is, the correlation is only considered valid if it exceeds the threshold τ_{sid} (here, τ_{sid} is empirically set to 0.5). The track T_κ with

$$\kappa = argmax_{k=1,\ldots,K}(corr(feat_I, k))$$

is then chosen for ID model adaptation.

The adaptation itself is made using a fixed learnrate α

$$ID^{k'} = \alpha idf + (1 - \alpha) ID^k$$

An alternative to using a learnrate-based adaptation would be, e.g., to update the model in a Bayesian fashion by taking the product of the modeled and observed densities

$$p(ID_t^k | z_t) \propto p(z_t | ID_{t-1}^k) p(ID_{t-1}^k | z_{t-1})$$

where the current model ID_t^k can be seen as the prior estimated from passed observations $\{z_1, \ldots, z_{t-1}\}$. The reason this is not done here is that this could quickly lead to a degenerate prior (a pdf with all but one value set to 0), such that errors in identification or in data association could not be recovered by later observations. Another reason is that sequential Bayesian estimation may not be the best suited method for this task in the first place, as one of its important assumptions is often not met: The independence of observations. An example is a scenario where visual identification is performed on one participant's face several times in sequence, at the same location, under the same conditions. It can be argued that these observations should not be considered independent, and that the whole sequence can be seen more or less as one observation. For these and other reasons, learnrate-based model adaptation (even using a variable learnrate) seems more appropriate and is commonly used in the literature.

In the explanations given above, visual and acoustic identity observations are treated uniformly and accumulated in a joint audio-visual ID model. As will

be shown later in Chapter 6, modeling acoustically and visually derived identities separately and fusing the results at the decision level may help increase identification performance. The reason is that speaker ID and face ID cues may be observed at quite different rates. Faces may in the best case be captured at framerate, or may be invisible for long periods of time. Voice identification is performed at best for 1 second segments and can only be made when the target person speaks. When updating one same model with both types of observations, the result may therefore be biased toward one modality based only on observation frequency. To avoid this, separate identity models ID_V^k and ID_A^k are kept for each track T_k and each model is adapted using only the features from the corresponding modality. The final audio-visual identity model ID^k (which is still used to compute correlations in the data association step) is then obtained as the product of modality-specific models. This is achieved by component-wise multiplication of the concerned pdfs, followed by renormalization of the result.

The fusion at decision level could also be done using the weighted sum of ID_V^k and ID_A^k. The choice of the product here is justified by the independence of the audio and visual modalities. Nevertheless, the weighted sum rule for modality fusion is also investigated and an evaluation of the two techniques as well as comparison to the single ID model method are given in Chapter 6.

4.2.1 Warping Confidence Values

In this section, a problem is addressed which is commonly encountered when combining the results of multiple classifiers: The normalization of confidence scores. Owing to the different nature of the classification algorithms, the derived confidences for face and voice identification, though bounded to the interval $[0, 1]$, are not directly comparable. Indeed, the actual confidence values are not evenly distributed between 0 and 1:

- The lowest possible confidence value γ_{min} is $1/n$, with n the number of hypotheses in the n-best list (Remember that the identification confidence is the normalized confidence score of the highest ranking identity in an n-best list output by the classifier).
- The highest confidence value γ_{max} is in principle 1, but for some classifier types actual confidence values are confined to a much smaller range (this depends e.g. also on the value n). For the speaker identification procedure described in Section 3.3.3, e.g., the difference between the top n confidences is usually very small, owing to the nature of the classifier, such that after renormalization of the n best confidences (with high enough n) the highest ranking confidence score is almost always very low.
- The optimum threshold value γ_{thresh} to distinguish known from unknown persons in the open set ID case consequently lies between γ_{min} and γ_{max}.

As a consequence, the derived confidence scores, as described in Sections 3.3.2 and 3.3.3 cannot be directly used to combine classifier outputs.

To remedy this, a warping function $f_{warp}(\cdot)$ is applied to the confidence scores. It is realized as a piecewise linear transformation, which maps the confidence values between γ_{min} and γ_{max} to the range $[\,0, 1]$ and maps the value γ_{thresh} to 0.5

$$f_{warp}(conf) = \begin{cases} \frac{1}{2} + \frac{1}{2}\frac{conf-\gamma_{thresh}}{\gamma_{max}-\gamma_{thresh}} & conf \geq \gamma_{thresh} \\ \frac{1}{2}\frac{conf-\gamma_{min}}{\gamma_{thresh}-\gamma_{min}} & else \end{cases}$$

The parameters γ_{min}, γ_{max} and γ_{thresh} can be chosen from the operating points of the individual classifiers, estimated, e.g., on experimental data. The resulting warped confidence values $f_{warp}(conf)$ are normalized to the same range and can therefore be used directly for comparison or fusion of identification results from the audio and visual modalities. Further, a common threshold can now be applied for the acceptance of known identities. In Chapter 6, we will see how the operating points for the here implemented face and voice recognizers are determined.

4.3 Joint Identity Filtering

While the previous sections were concerned with the tracking of persons and their identification on an individual basis, the topic of this section is the tracking of *Identities* in the smart space. This means that the presence of known persons in the space is to be detected, their locations found if possible, and their identities jointly derived. Indeed, estimating identities individually for tracks can lead to unwanted results, such as wrongfully assigning the same identity to two or more tracks. In the presence of ambiguous measurements, deciding on which identity to assign to which tracked person, on which person to reject as unknown in the case of conflict, etc., on a global scale can be a non-trivial matter. The solution proposed here is a probabilistic filtering approach, that jointly estimates the probabilities for the presence, location and identity of multiple known persons. The algorithm, referred to as the *Joint Identity Tracking (JIT)* filter, is described in the following.

4.3.1 Assumptions and Task Definition

Two main assumptions are made at this point, that will guide the design of the identity tracking filter:

- Information about tracks is inherently flawed. The results of multiple person tracking in natural, cluttered environments, cannot be expected to be flawless. False tracks may be generated, persons may not be detected at all, estimated locations may be erroneous, tracks may be swapped, etc. Especially in the last case, observations that have been accumulated for one target person are after the swap wrongfully assigned to another person. The identity tracking system must be able to recover from such errors.
- Identification results are inherently flawed. For both modalities, single identification results, no matter how high the estimated confidence, are not absolutely trustworthy. This means that observations for one person should be accumulated to increase the accuracy of the result. The constancy, but also the frequency of accumulated results can serve as indicators for the quality of the identification, especially when unknown persons or unknown sources of noise are present.

Given these assumptions, the number of persons present in the smart space, e.g., cannot be reliably estimated using the results of tracking or the number of recognized identities alone, but should be derived jointly using both sources of information.

Let us now define the relevant terms used in the description of the identity tracking task:

- Identities: Here, the term is used in a simplifying way to designate only "known" identities, i.e. those for which classifiers have been trained beforehand. The identities of unknown persons are not considered, i.e. the task does not include differentiating between unknown persons (this is the object of general person tracking). The set of identities is denoted $\mathcal{I} = \{id_1, id_2 \dots id_{N_{ids}}\}$. Generally, only a small subset of the known identities is present in the smart space at one time.
- Tracks: These are entities which have been detected in the smart space and the locations of which are being tracked. Tracks may or may not represent actual persons. False tracks are often enough generated, caused by false detections, noisy measurements or estimation errors.
- Persons: The word is used in its common meaning to designate actual persons evolving in the space. The number of persons present in the space is variable. They are further divided into known "focus" persons and unknown persons. Persons are not necessarily tracked. "Tracked persons" are those whose location coincides with a track hypothesis. This is in opposition to
- Hidden persons: Those are persons which are found present in the space, though their location can not be estimated. The presence of such persons

can be detected, e.g., by the highly confident or repeated recognition of an identity in the space, for which no track information is available. This may be because the corresponding track got lost, or could not be initialized in the first place.

The task of identity tracking consists in detecting the presence of known persons (identities) in the smart space and determining their location. An important part of the task is is also *not to detect* identities which are absent from the space, due to wrongful recognition of unknown persons. Any of these subtasks may be accomplished with a more or less high degree of accuracy.

The joint identity tracking filter realizes this by managing a set of person models $\mathcal{P} = \{P_1, \ldots, P_{N_P}\}$ for the persons present in the space. This set is composed of tracked persons $\mathcal{P}_\mathcal{T}$ and of "hidden persons" $\mathcal{P}_\mathcal{H}$

$$\mathcal{P} = \mathcal{P}_\mathcal{T} \cup \mathcal{P}_\mathcal{H}$$

Let I be a random variable defined over the set of identities $\{id_1, id_2 \ldots id_{N_{ids}}\}$ and L be a random variable representing their possible locations. An identity can be hypothesized as being associated to one of the tracked persons $\mathcal{P}_\mathcal{T}$, to one of the hidden persons $\mathcal{P}_\mathcal{H}$ or to be absent from the smart space. L is therefore defined over the set $\{P_0, P_1, \ldots, P_{N_P}\}$ with P_0 representing the hypothesis that an identity is outside of the smart space. Based on the speaker and face ID cues observed so far, the filter approximates the joint probability $p(I, L|z_{1:t})$. The corresponding marginal distributions $p(I|L, z_{1:t})$ and $p(L|I, z_{1:t})$) are of special interest. For $n = 1, \ldots, N_P$, $p(I|L = n, z_{1:t})$ represents the belief in the identity of person P_n. In the case of individual estimation for tracked persons, it is equivalent to the probability density function ID^k modeled in the corresponding track, as described in Section 4.2. For $i = id_1, \ldots, id_n$, $p(L|I = i, z_{1:t}))$ represents the belief in the location of identity id_i. It is a discrete probability function over $\{P_0, P_1, \ldots, P_{N_P}\}$. It abstracts from actual spatial coordinates in the scene by assigning identities to tracked locations, unknown locations inside the smart space, or undefined locations outside of the smart space. The next section will show how the joint probability distribution $p(I, L|z_{1:t})$ is approximated from the marginal distributions in an iterative estimation process.

4.3.2 Joint Update in Identity Space

When estimating the identities of multiple persons, two conditions must always be satisfied:

- One person can be associated at most one identity. This is common sense and is translated in probabilistic terms by requiring that the probability distribution modeling the identity of a person normalizes to one.

$$\sum_{i=1}^{N_{ids}} p(I = i|L = n, z_{1:t}) = 1,\ n = 1, \ldots, N_P$$

- Two persons cannot share a same identity or, expressed in other terms, one identity cannot be hypothesized to be present in two places at the same time. This means that the distribution modeling the location of an identity must also normalize to one.

$$\sum_{n=0}^{N_P} p(L = n|I = i, z_{1:t}) = 1,\ i = 1, \ldots, N_{ids}$$

The problem consists in keeping both constraints met when updating the joint probability $p(I, L|z_{1:t-1})$ given a new observation z_t. This is done in an iterative refinement process, as follows:

1. Update: Each time a new observation is associated to a modeled person P_n, the marginal probability for that person is first updated. Let *idf* be the discrete pdf modeling the observed identity. The update is then made as described in Section 4.2 as

$$p(I|L = n, z_{1:t}) = \alpha idf + (1 - \alpha)p(I|L = n, z_{1:t-1})$$

   with a fixed learnrate α. Since both distributions were normalized density functions, the result is also a normalized pdf with values summing up to 1.

2. Location normalization: Since modifying the belief about the identity of a person also changes the belief about the locations of all identities modeled by that person, the marginal distributions for identity locations are renormalized

$$p(L|I = i, z_{1:t}) = \frac{p(L|I = i, z_{1:t-1})}{\sum_{n=0}^{N_P} p(L = n|I = i, z_{1:t-1})},\ i = 1, \ldots, N_{ids}$$

3. Identity normalization: Modifying the belief about locations in turn changes the belief about the identity of each person. The marginal distributions for identities are therefore renormalized

$$p'(I|L = n, z_{1:t}) = \frac{p(I|L = n, z_{1:t})}{\sum_{i=1}^{N_{ids}} p(I = i|L = n, z_{1:t})},\ n = 1, \ldots, N_P$$

   Note that no similar renormalization is made for $n = 0$, as it is not contradictory for any number of identities to be in an undefined location *outside* of the space.

4. Steps 2 and 3 are repeated until the cumulative difference between values of the marginal distributions from one iteration to the next falls below a specified threshold T_{break} or a maximum number of iterations N_{break} is reached. Experiments have shown that a small number of iterations (generally less than 5) is sufficient to achieve convergence.

After the update step, the value of $p(I = i, L = n|z_{1:t})$ represents the probability, given the observation sequence, that identity id_i is present in the space, represented by person P_n, or absent from the space (for $n = 0$).

The output of the filter is the most likely identity for each of the modeled persons, as well as the confidence in this identity. This is done for each person P_η by finding the identity ι_η which maximizes both marginal distributions: $\iota_\eta = argmax_{i=1}^{N_{ids}} p(I = i, L = \eta)$ and $\eta = argmax_{n=1}^{N_P} p(I = \iota_\eta, L = n)$. First, the likeliest identity for person P_η is determined. If the found identity does not represent a maximum of $p(L|I)$ (i.e. the likelihood for assigning it to another person is higher), the second best identity is taken, and so forth. If no identity can be assigned to P_η, it is considered an unknown person ($\iota_\eta = id_0$). In the maximum search process, priority is given to tracked persons: If the same maximum i is found for a tracked person P_n and a hidden person P_m, the tracked person is chosen for output and the confidence is taken as the sum of confidences $p(I = i, L = n) + p(I = i, L = m)$. In this case, it is likely P_m was wrongfully modeled as a separate person and rather expresses an uncertainty in the location of P_n. This can happen when localized or non-localized ID features cannot be mapped directly to tracked persons, as will be explained in the next section.

In this manner, only one identity is output per person, and no two persons are assigned the same identity. The output of the filter is therefore the set of assignments $\left\{(P_\eta, id_{\iota_\eta}, conf(\eta, \iota_\eta))\right\}_{\eta=1}^{N_P}$, with $conf(\eta, \iota_\eta) = p(I = \iota_\eta, L = \eta)$.

4.3.3 Person Creation, Deletion and Data Association

This section explains the process of creating and deleting person models, as compared to the creation and deletion of tracks, described in Section 4.1.8.

Person models are created based on two criteria. The first is tracking information. When a new track is initialized and validated, as described in Section 4.1.8, a corresponding person model is automatically created and added to the set of tracked persons $\mathcal{P}_\mathcal{T}$. The second is information about recognized identities. When localized ID features are extracted, they are mapped to tracks in a probabilistic data association process, as described in Section 4.2. If no matching track is found, however, the observation is hypothesized to originate from a person present in the smart space for which a track could not be initialized. Therefore, an association to the set of hidden persons $\mathcal{P}_\mathcal{H}$ is attempted. If this

association also fails, a new person model is created and added to $\mathcal{P}_\mathcal{H}$, and its identity pdf updated based on the observation.

While the association to tracked persons is made using both identity correlation and spatial coverage, the association to hidden persons is made based only on identity correlation. For this, the marginal distribution $p(I|L = n)$ for a hidden person P_n is taken and compared to the density idf given in the observation. As in Section 4.2, the Bhattacharyya coefficient is used as measure to compute a correlation coefficient

$$Id^n_{corr} = B_C(idf, p(I|L = n)).$$

The correlation to a uniform pdf over the space of identities $Id^0_{corr} = B_C(idf, U)$ is taken as a threshold, such that association to P_n is only made if $Id^n_{corr} > Id^0_{corr}$.

For the deletion of person models, two cases must again be considered. The first is when a person track is lost, as explained in Section 4.1.8. The corresponding person model is then not immediately deleted, but moved to the set of hidden persons $\mathcal{P}_\mathcal{H}$: It is hypothesized that the person may still be present, though not tracked. The question is then when to delete models for hidden persons, which is the second case. This is done similarly as for the deletion of identified tracks: Models for hidden persons are kept alive as long as they are regularly observed. Models for persons hypothesized as unknown (for which the accumulated identification confidence does not exceed a specified threshold) are deleted if no observation can be associated to them for a certain period of time $\tau_{sustain}$. Models for persons hypothesized as known are deleted only after a longer period τ_{known}. The reason is, again, that they are much less likely to be false hypotheses caused by data association errors, errors in recognition, etc.

4.3.4 Evaluating Identities

Although the proposed Joint Identity Tracking Filter estimates non-parametric probability distributions for locations and identities, the outputs of the algorithm need to be discretized for comparison to the ground truth in performance evaluations (see Chapter 5). This means that one discrete identity and one definite 3D position in the scene have to be computed for each hypothesized person. Section 4.3.2 already described how marginal distributions are used to derive a set of output assignments $\{(P_n, id_{\iota_n}, conf(n, \iota_n))\}_{n=1}^{N_P}$. The output location vector $\vec{x}_n$ for tracked persons is then taken as the location of the corresponding track. The location vector $\vec{x}_n$ is invalidated for hidden persons. Although the localization subtask is considered failed for these persons, the identification performance can still be evaluated.

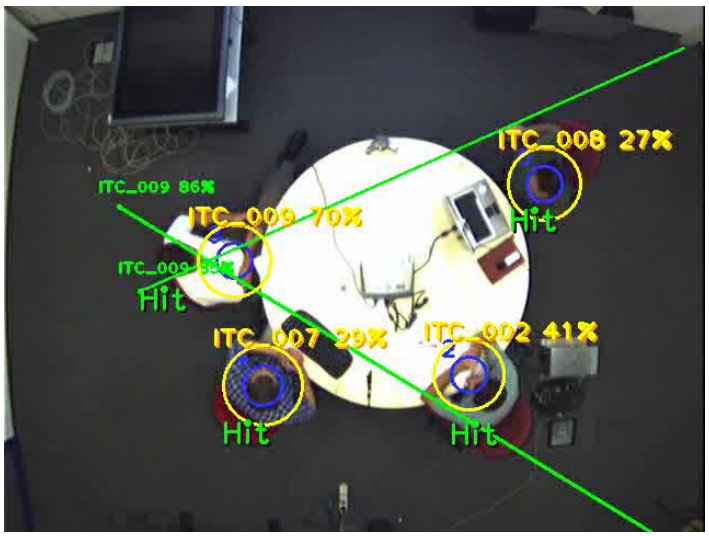


Figure 4.7: The output of the integrated identity tracking system. The blue and yellow circles represent the person tracker hypotheses and the person models, respectively. The identities for recognized persons are printed on top of the respective models. The green lines indicate face identification hits for the current frame, in this example made by two of the four corner cameras for one room occupant.

Finally, the confidence $conf(n, \iota_n)$ is used to decide if the considered person is known or unknown. As explained in Section 3.3, the open set identification task is realized by thresholding the identification confidence values. If $conf(n, \iota_n) < Th_{known}$, the person is considered unknown and the index id_0 is output. Otherwise the person is considered known and id_{ι_n} is output. Different values for the acceptance threshold Th_{known} will be investigated in Chapter 6. In summary, the output of the JIT filter is, for every time point t, a set of hypothesis identities and locations

$$\{(id_{\iota_n}, \vec{x}_n)\}_{n=1}^{N_P}$$

and the confidence information is not further used.

Figure 4.7 shows an example output of the integrated identity tracking system on a CLEAR seminar with four known participants.

5 Performance Metrics

Performance evaluations are an important factor in in the development of algorithms an techniques. Systematic evaluations using large benchmark datasets and producing quantitative, statistically significant results are essential to compare the strengths and weaknesses of different techniques - and measure progress. The task of multiple identity tracking with distantly placed sensors, as defined in this thesis, is a new problem definition, which has, to my knowledge, not been addressed to such extent and detail in the literature. As a consequence, the problem of defining metrics to measure the performance of identity tracking systems and methods has also not been solved satisfactorily. Perhaps the main reason for the lack of commonly agreed on metrics and evaluation procedures so far is that large benchmark databases of multimodal, multi-sensory inputs, recorded in realistic conditions for natural multiple person interaction scenarios were, until the past few years, just not available. The CLEAR evaluation workshops [106; 105] have changed this situation by providing to the scientific community databases featuring hundreds of hours of multimodal recordings of natural meeting and seminar scenarios, captured under realistic and quite varied conditions. The evaluations already provided a forum for the evaluation of novel multiple person tracking tasks, and some of the metrics used in these evaluations were developed in the course of this thesis to overcome the lack of standardized evaluation procedures so far. Here, a new set of metrics is also presented, complementing the previous ones, for the evaluation of algorithms and systems performing simultaneous tracking and identification. As a consequence, two sets of performance evaluation metrics are hereby introduced:

- Multiple Object Tracking (MOT) Metrics: These are used to measure the accuracy of person tracking. This includes measuring if the number of persons present could be correctly determined, if their locations are correctly estimated, how precise the localization is done, how well track integrity is kept, etc.
- Multiple Identity Tracking (MIT) Metrics: These metrics measure the accuracy of identity tracking. The objects of evaluation here are exclusively *known identities*, i.e. the identities of known persons. They measure if the presence of known identities in the smart space is correctly detected, if identities are correctly determined and how well their locations are estimated.

In addition, the well known measures of correct classification, false classification, false rejection and false acceptance (as explained in Section 3.3), commonly used e.g. in the field of biometric verification, are used here to evaluate specifically open set identification performance. These measures can however not be applied or interpreted in the standard way, which is why modifications to the standard evaluation procedure are explained in Section 5.3.

5.1 Performance Metrics for Multiple Object Tracking

In the course of this thesis, a novel method to systematically evaluate the performance of multiple object trackers has been developed. A procedure to detect the basic types of errors produced by multiple object trackers is presented and two novel metrics are introduced, the Multiple Object Tracking Precision ($MOTP$), and the Multiple Object Tracking Accuracy($MOTA$), that intuitively express a tracker's overall strengths and are suitable for use in large-scale performance evaluations.

To allow a better understanding of the proposed metrics, the qualities we expect from an ideal multiple object tracker are first explained: It should at all points in time find the correct number of objects present and estimate the position of each object as precisely as possible (Note that properties such as the contour, orientation or speed of objects are not explicitly considered here). It should also keep consistent track of each object over time: Each object should be assigned a unique track ID which stays constant throughout the tracking sequence (even after temporary occlusion, etc). This leads to the following design criteria for performance metrics:

- They should allow to judge a tracker's precision in determining exact object locations.
- They should reflect its ability to consistently track object configurations through time, i.e. to correctly trace object trajectories, producing exactly one trajectory per object.

Additionally, we expect useful metrics

- to have as few free parameters, adjustable thresholds, etc, as possible to help make evaluations straightforward and keep results comparable.
- to be clear, easily understandable and behave according to human intuition, especially in the occurrence of multiple errors of different types or of uneven repartition of errors throughout the sequence.

- to be general enough to allow comparison of most types of trackers (2D, 3D trackers, object centroid trackers or object extension trackers, visual or acoustic trackers, etc).
- to be few in number and yet expressive, so that they may be used, e.g., in large-scale evaluations where many systems need to be compared.

Based on the above criteria, a procedure is proposed for the systematic and objective evaluation of a tracker's characteristics. Assuming that for every time frame t a multiple object tracker outputs a set of hypotheses $\{h_1 \dots h_m\}$ for a set of visible objects $\{o_1 \dots o_n\}$, the evaluation procedure comprises the following steps:
For each time frame t,

- Establish the best possible correspondence between hypotheses h_j and objects o_i
- For each found correspondence, compute the error in the object's location estimation.
- Accumulate all correspondence errors:
  - Count all objects for which no hypothesis was output as misses.
  - Count all tracker hypotheses for which no real object exists as false positives.
  - Count all occurrences where the tracking hypothesis for an object changed compared to previous frames as mismatch errors. This could happen, e.g., when two or more objects are swapped as they pass close to each other, or when an object track is reinitialized with a different track ID, after it was previously lost because of occlusion.

Then, the tracking performance can be intuitively expressed in two numbers: the "tracking precision", which expresses how well exact person locations are estimated, and the "tracking accuracy", which shows how many mistakes the tracker made in terms of misses, false positives, mismatches, failures to recover tracks, etc. These measures will be explained in detail in the latter part of this section.

5.1.1 Establishing Correspondences Between Objects and Tracker Hypotheses

As explained above, the first step in evaluating the performance of a multiple object tracker is finding a continuous mapping between the sequence of object hypotheses $\{h_1 \dots h_m\}$ output by the tracker in each frame and the real target objects $\{o_1 \dots o_n\}$. This is illustrated in Fig. 5.1. Naively, one would match

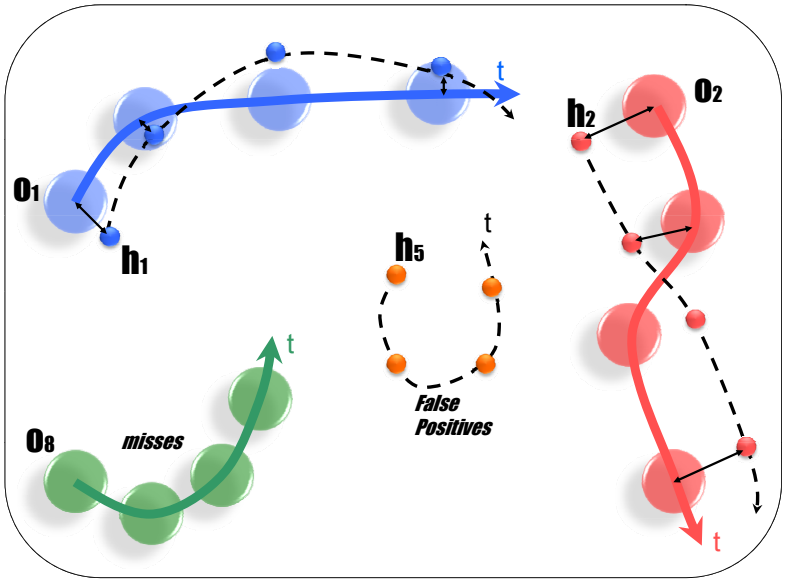


Figure 5.1: Mapping tracker hypotheses to objects. In the easiest case, matching the closest object-hypothesis pairs for each time frame t is sufficient.

the closest object-hypothesis pairs and treat all remaining objects as misses and all remaining hypotheses as false positives. A few important points need to be considered, though, which make the procedure less straightforward.

Valid Correspondences

First of all, the correspondence between an object o_i and a hypothesis h_j should not be made if their distance $dist_{i,j}$ exceeds a certain threshold Th_{dist}. There is a certain conceptual boundary beyond which we can no longer speak of an error in position estimation, but should rather argue that the tracker has missed the object and is tracking something else. This is illustrated in Fig. 5.2(a). For object extension trackers (i.e. trackers that also estimate the size of objects or the area occupied by them), distance could be expressed in terms of the overlap between object and hypothesis and the threshold Th_{dist} could be set to zero overlap. For object centroid trackers, one could simply use the Euclidian distance, in 2D image coordinates or in real 3D world coordinates, between object and hypothesis centers, and the threshold could be, e.g., the average width of a person in pixels or cm. The optimal setting for Th_{dist} therefore depends on the application task, the size of objects involved and the distance measure used, and cannot be defined for the general case . In the following, correspondences will be referred to as *valid* if $dist_{i,j} < Th_{dist}$.

Consistent Tracking

Second, to measure a tracker's ability to label objects consistently, one has to detect when conflicting correspondences have been made for a target object over time. Fig. 5.2(b) illustrates the problem. Here, one track was mistakenly assigned to 3 different objects over the course of time. A mismatch can occur when objects come close to each other and the tracker wrongfully swaps their tracks. It can also occur when a track was lost and reinitialized with a different track index. One way to measure such errors could be to decide on a "best" mapping (o_i, h_j) for every object o_i and hypothesis h_j, e.g. based on the initial correspondence made for o_i, or the correspondence (o_i, h_j) most frequently made in the whole sequence. One would then count all correspondences where this mapping is violated as errors. In some cases, this kind of measure can however become non-intuitive. As shown in Fig. 5.2(c), if, for example, the identity of object o_i is swapped just once in the course of the tracking sequence, the time point at which the swap occurs drastically influences the value output by such an error measure.

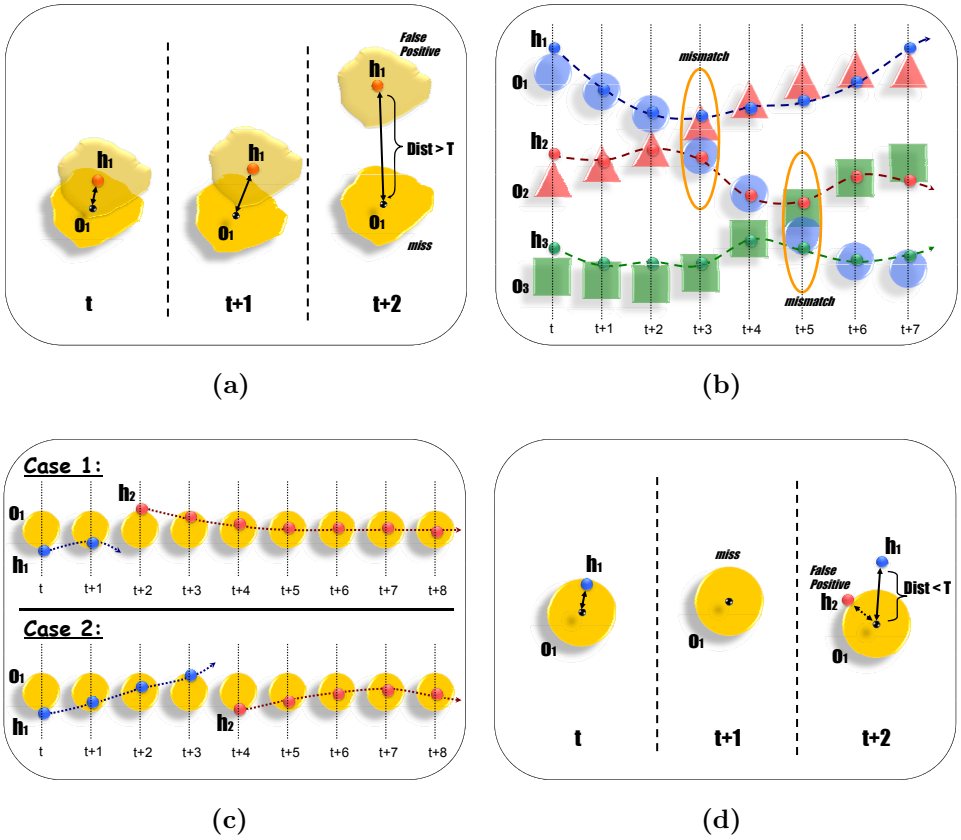
This is why a different approach is followed here: only count mismatch errors once at the time frames where a change in object-hypothesis mappings is made and consider the correspondences in intermediate segments as correct. Especially in cases where many objects are being tracked and mismatches are frequent, this gives us a more intuitive and expressive error measure. To detect when a mismatch error occurs, a list of object-hypothesis mappings is constructed. Let $M_t = \{(o_i, h_j)\}$ be the set of mappings made up to time t and let $M_0 = \{\}$. Then, if a new correspondence is made at time $t + 1$ between o_i and h_k which contradicts a mapping (o_i, h_j) in M_t, a mismatch error is counted and (o_i, h_j) is replaced by (o_i, h_k) in M_{t+1}.

The so constructed mapping function M_t can now help to establish optimal correspondences between objects and hypotheses at time $t+1$, in the case multiple valid choices exist. Fig. 5.2(d) shows such a case. When it is not clear, which hypothesis to match to an object o_i, priority is given to h_o with $(o_i, h_o) \in M_t$, as this is most likely the correct track. Other hypotheses are considered false positives, and could have occurred because the tracker output several hypotheses for o_i, or because a hypothesis that previously tracked another object accidentally crossed over to o_i.

Mapping Procedure

Having clarified all the design choices behind our strategy for constructing object-hypothesis correspondences, we summarize the procedure:

Let $M_0 = \{\}$. For every time frame t,



(a) (b)

(c) (d)

Figure 5.2: Optimal correspondences and error measures. Fig. 5.2(a): When the distance between o_1 and h_1 exceeds a certain threshold Th_{dist}, one can no longer make a correspondence. Instead, o_1 is considered missed and h_1 becomes a false positive. Fig. 5.2(b): Mismatched tracks. Here, h_2 is first mapped to o_2. After a few frames, though, o_1 and o_2 cross paths and h_2 follows the wrong object. Later, it wrongfully swaps again to o_3. Fig. 5.2(c): Problems when using a sequence-level "best" object-hypothesis mapping based on most frequently made correspondences. In the first case, o_1 is tracked just 2 frames by h_1, before the track is taken over by h_2. In the second case, h_1 tracks o_1 for almost half of the sequence. In both cases, a "best" mapping would pair h_2 and o_1. This, however, leads to counting 2 mismatch errors for *case 1* and 4 errors for *case 2*, although in both cases only one error of the same kind was made. Fig. 5.2(d): Correct reinitialization of a track. At time t, o_1 is tracked by h_1. At $t+1$, the track is lost. At $t+2$, two valid hypotheses exist. The correspondence is made with h_1 although h_2 is closer to o_1, based on the knowledge of previous mappings up to time $t+1$.

1. For every mapping (o_i, h_j) in M_{t-1}, verify if it is still valid. If object o_i is still visible and tracker hypothesis h_j still exists at time t, and if their distance does not exceed the threshold Th_{dist}, make the correspondence between o_i and h_j for time frame t.

2. For all objects for which no correspondence was made yet, try to find a matching hypothesis. Allow only one to one matches, and pairs for which the distance does not exceed Th_{dist}. The matching should be made in a way that minimizes the total object-hypothesis distance error for the concerned objects. This is a minimum weight assignment problem, and is solved using Munkres' algorithm [82] with polynomial computational complexity. If a correspondence (o_i, h_k) is made that contradicts a mapping (o_i, h_j) in M_{t-1}, replace (o_i, h_j) with (o_i, h_k) in M_t. Count this as a mismatch error and let mme_t be the number of mismatch errors for time frame t.

3. After the first two steps, a complete set of matching pairs for the current time frame is known. Let c_t be the number of matches found for frame t. For each of theses matches, calculate the distance d_t^i between the object o_i and its corresponding hypothesis.

4. All remaining hypotheses are considered false positives. Similarly, all remaining objects are considered misses. Let fp_t and m_t be the number of false positives and misses respectively for frame t. Let also g_t be the number of target objects present at time frame t.

5. Repeat the procedure from step 1 for the next time frame. Note that since for the initial frame, the set of mappings M_0 is empty, all correspondences made are initial and no mismatch errors occur.

In this way, a continuous mapping between objects and tracker hypotheses is defined and all tracking errors are accounted for.

5.1.2 MOT Metrics

Based on the matching strategy described above, two very intuitive measures can be defined.

1. The *M*ultiple *O*bject *T*racking *P*recision ($MOTP$).

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}$$

   It is the total localization error for matched object-hypothesis pairs over all time frames, averaged by the total number of matches made. It shows the ability of the tracker to estimate precise object positions, independent

of its skill at recognizing object configurations, keeping consistent trajectories, etc.

2. The *M*ultiple *O*bject *T*racking *A*ccuracy ($MOTA$).

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}$$

where m_t, fp_t and mme_t represent the number of misses, of false positives and of mismatches, respectively, for time frame t. The $MOTA$ can be seen as derived from 3 error rates:

$$\overline{m} = \frac{\sum_t m_t}{\sum_t g_t},$$

the rate of misses in the sequence, computed over the total number of objects present in all frames,

$$\overline{fp} = \frac{\sum_t fp_t}{\sum_t g_t},$$

the rate of false positives, and

$$\overline{mme} = \frac{\sum_t mme_t}{\sum_t g_t},$$

the rate of mismatches.

Summing up over the different error rates gives us the total error rate E_{tot}, and $1 - E_{tot}$ is the resulting tracking accuracy. The $MOTA$ accounts for all object configuration errors made by the tracker, false positives, misses, mismatches, over all frames. It is similar to metrics widely used in other domains (such as the Word Error Rate (WER), commonly used in speech recognition) and gives a very intuitive measure of the tracker's performance at detecting objects and updating their trajectories, independent of the precision with which the object locations are estimated.

Remark on Computing Averages: Note that for both $MOTP$ and $MOTA$, it is important to first sum up all errors across frames before a final average or ratio can be computed. The reason is that computing ratios r_t for each frame t independently before calculating a global average $\frac{1}{n}\sum_t r_t$ for all n frames (such as, e.g., for the $\overline{FP}$ and $\overline{FN}$ measures in [97]), can lead to non-intuitive results. This is illustrated in Fig. 5.3. Although the tracker consistently missed most objects in the sequence, computing ratios independently per frame and then averaging would still yield only 50% miss rate. Summing up all misses first and computing a single global ratio, on the other hand, produces a more intuitive result of 80% miss rate.

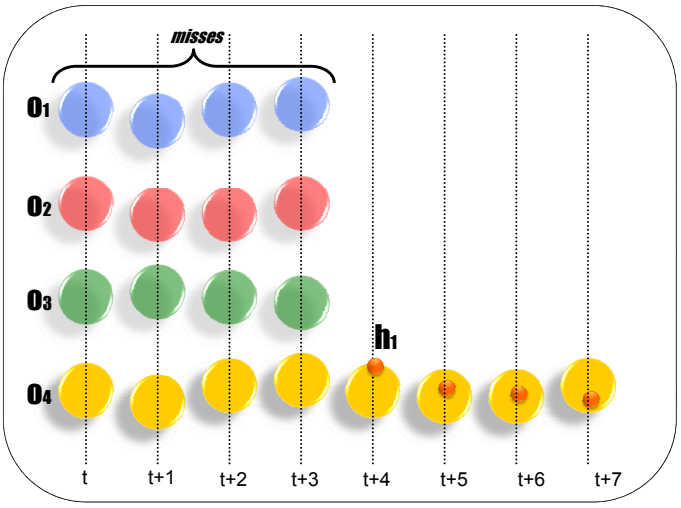
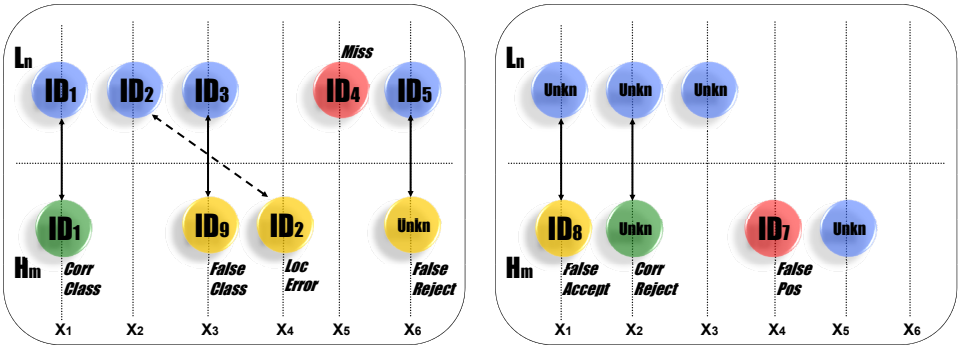


Figure 5.3: Computing error ratios. Assume a sequence length of 8 frames. For frames t_1 to t_4, 4 objects $o_1 \ldots o_4$ are visible, but none is being tracked. For frames t_5 to t_8, only o_4 remains visible, and is being consistently tracked by h_1. In each frame $t_1 \ldots t_4$, 4 objects are missed, resulting in 100% miss rate. In each frame $t_5 \ldots t_8$, the miss rate is 0%. Averaging these frame level error rates yields a global result of $\frac{1}{8}(4 \cdot 100 + 4 \cdot 0) = 50\%$ miss rate. On the other hand, summing up all errors first, and computing a global ratio yields a far more intuitive result of $16 misses / 20 objects = 80\%$.



(a) (b)

Figure 5.4: The different types of errors that can be produced in identity tracking. The top rows represent the ground truth an the bottom rows the tracker hypothesis. The x-axis represents a one-dimensional abstraction of locations in the smart space. In the case the ground truth person represents a known identity (5.4(a)), four types of errors are possible: false localization, false classification, false rejection and a complete miss. Only if both the location and the identity are correctly inferred is no error made. Note tha in the case of a pure *localization* error, a correct classification is still counted. Likewise, incorrectly identified persons can still be correctly localized. 5.4(b) shows the cases where the ground truth represents an unknown person. Two types of errors can be made: a false acceptance and a false positive hypothesis. In the case the unknown person is tracked and correctly recognized as such, no error occurs. All other cases are ignored in the evaluation of identity tracking.

5.2 Performance Metrics for Identity Tracking

The goal of identity tracking is to recognize and localize known focus persons interacting with several unknown ones in a smart environment. Similarly to the *MOT* metrics, we first count the different types of errors made by the identity tracker in an observation sequence, and then define appropriate error measures. Possible errors include failing to determine the correct identity of a known person, failing to correctly estimate his or her location, falsely rejecting a known person as unknown, falsely recognizing an unknown person as known, failing to detect the presence of a known identity in the smart space and falsely hypothesizing the presence of a known person which is absent. Figure 5.4 shows an illustration of the different error types. The evaluation procedure is now defined in accordance:

Let $\mathcal{P} = \{P_1, \ldots, P_{N_P}\}$ be the set of persons present in smart space at time t, divided into focus persons $\mathcal{P}_\mathcal{F}$ and unknown persons $\mathcal{P}_\mathcal{U}$. Let $L = \{l_1, \ldots, l_{N_P}\}$ be the set of ground truth person labels and $H = \{h_1, \ldots, h_{N_H}\}$ the hypothesis

output by the identity tracker for time t. Each label and each tracker hypothesis consist of an identity $\iota \in \{id_0, id_1, \ldots, id_{N_{ids}}\}$ and a location vector $\vec{x}$, as defined in Section 4.3.4:

$$l_n = (\iota_n, \vec{x}_n)$$

$$h_m = (\iota_m, \vec{x}_m)$$

with id_0 used to denote that the person in question is unknown and $\vec{x} = \vec{0}$ used to denote that his or her location is undefined. For the purposes of identity tracking, an identity is considered correctly localized as long as the distance error to the ground truth location is below a certain threshold. Here, the same threshold as for the MOT metrics, Th_{dist} is used. The evaluation of fine localization accuracy is thus considered part of the person tracking task, and is not repeated here.

For each time frame t:

1. For all labeled focus persons $l_n = (\iota_n, \vec{x}_n)$, with $\iota_n \neq id_0$, try to find a corresponding hypothesis h_m. Four cases may arise:

   - Perfect match: If h_m can be found with $\iota_m = \iota_n$ and $dist(\vec{x}_m, \vec{x}_n) < Th_{dist}$, count this as a correct classification.
   - Location mismatch: If h_m is found with $\iota_m = \iota_n$ and $dist(\vec{x}_m, \vec{x}_n) > Th_{dist}$, count this also as a correct classification, but additionally count a localization error.
   - Identity mismatch: If, on the contrary, h_m is found with matching location, but differing identity, consider this an identification error. The error is further refined based on the hypothesized identity. If $\iota_m = id_0$, count a false rejection error. Otherwise, count the error as a false classification.

   Let CC_t, LE_t, FR_t and FC_t, be the number of correct classifications, localization errors, false rejections and false classifications made for time frame t. Let also MS_t, be the number of identities missed in frame t.

2. For all labeled unknown persons $l_n = (\iota_n, \vec{x}_n)$, with $\iota_n = id_0$, try to find a corresponding hypothesis h_m, which has not been matched before based on spatial proximity. If h_m can be found with $dist(\vec{x}_m, \vec{x}_n) < Th_{dist}$, the decision is made based on the hypothesized identity. If $\iota_m = id_0$, count a correct rejection (this is not an error, but will be used later in ratio computations). Otherwise, count a false acceptance error. If no matching hypothesis can be found, this is not considered an *identity tracking* error, as no known identity is involved. Let CR_t and FA_t be the number of correct rejections and false acceptances made for time frame t.

3. For all remaining unmatched hypotheses h_m, decide based on their identity. If $\iota_m = id_0$, ignore the hypothesis, as it again does not constitute an error in the sense of *identity* tracking (although it is an error under the aspect of person tracking and is counted in the $MOTA$ score). If $\iota_m \neq id_0$, count this as a false positive error and let FP_t be the number of false positives found for time frame t.

4. Repeat the procedure from step 1 for time frame $t + 1$.

In this way, all possible error types are accumulated on a frame basis. Let GT_t be the total number of labeled persons and GTF_t the number of labeled focus persons for time t. The accumulated, sequence level scores, GT, GTF, CC, FC, FR, FA, CR, MS, FP, are then accumulated by summing up over corresponding frame-level scores.

5.2.1 Multiple Identity Tracking Accuracy

From the above computed sequence-level scores, the following metrics for identity tracking performance are computed:

- the *L*ocalization *A*ccuracy (LA).

$$LA = 1 - \frac{LE + MS}{GTF}$$

  It measures the performance of the tracker at detecting and localizing known identities. Swapping identities, hypothesizing them at the wrong place, as well as missing identities completely (hypothesizing them to be outside of the smart space) are all considered localization errors.

- The *I*dentification *A*ccuracy (IA).

$$IA = 1 - \frac{MS + FR + FC + FA + FP}{GT}$$

  It measures the accuracy of the tracker at detecting and recognizing known identities in the presence of unknown persons. Failing to recognize identities (either by missing them completely or by rejecting them as unknown), misclassifying them, as well as falsely stating their presence (either by falsely recognizing unknown persons or by hypothesizing additional known persons) are penalized.

- The *M*ultiple *I*dentity *T*racking *A*ccuracy ($MITA$).

$$MITA = \frac{LA + IA}{2}$$

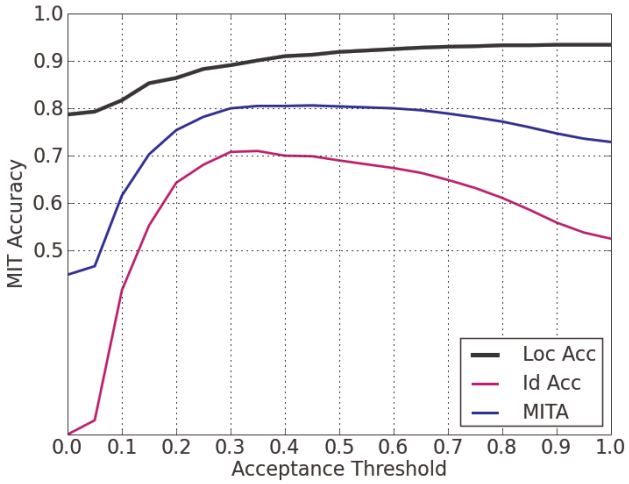


Figure 5.5: Example curves for localization accuracy (LA), identification accuracy (IA) and Multiple Identity Tracking Accuracy ($MITA$). The values of the measures are plotted as a function of the acceptance threshold for identification in the open set case. The $MITA$ represents the average of teh two other scores.

It is calculated over the entire observation sequence and taken as the average of the localization accuracy and the identification accuracy. Of course, a weighted average could also be computed instead. As both localization and identification are considered equally important here, though, equal weights are taken. Similarly to the $MOTA$, the $MITA$ allows a quick assessment of the overall identity tracking performance.

As many of the scores used in the computation of these metrics (CCR, FCR, FRR, FAR) are dependent on the acceptance threshold Th_{known} for open set identification, the LA, IA and $MITA$ are visualized as curves over the range of possible values of Th_{known}. The highest point on these curves represents the best achievable accuracy rate, assuming equal importance of individual errors (similar to the EER, as described in Section 3.3). Examples of LA, IA and $MITA$ curves can be seen in Fig. 5.5.

5.3 Evaluating Open Set Identification Performance

Open set identification performance is usually viewed independently of detection and localization performance. In the fields of multimodal identification or biometric verification, commonly used measures to evaluate open set classifiers

are the CCR, FCR, FRR and FAR (see Section 3.3). To provide an easier overview of these error rates as the threshold for acceptance of test subjects is varied, they are usually visualized jointly in Receiver Operating Characteristic (ROC) plots. Such ROC plots will also be used here to visualize some aspects of open set identification performance in the case of multiple identity tracking. Some important differences to the standard evaluation procedure exist, though, which should be clarified here: ROC curves are usually employed to evaluate identification or verification tasks, where the detection of the subjects to be identified plays no role. This is, e.g., the case in image-based face identification, where a set of test samples containing (genuine and impostor) faces is presented to a classifier, which then needs to decide only on the identity of the pre-selected samples. The evaluation conditions are somewhat different in the case of identity tracking, as defined here, as correctly identifying a person, e.g., also requires to detect the person in the first place. The $MITA$ metric is designed to reflect this fact by including misses and false positives in the error computation.

The CCR, FCR, FRR and FAR measures, on the other hand, are computed for each person using only the subset of frames where the person is detected (where a corresponding hypothesis is produced). This means that detection errors directly influence the number of samples presented for identification. An increase of the miss rate can, e.g., cause a sharp improvement of the correct classification rate if the set of persons (or frames) which are no longer tracked coincides with the set of persons (or frames) which are hard to identify, such that the CCR is computed only for a few well observable test cases. This is why the ROC curves for open set identification performance are used only as additional diagnostics and should always be interpreted in conjunction with the corresponding MIT scores.

Another difference is that the CCR, FCR, FRR and FAR measures are here computed by averaging accumulated frame level scores, although the output of the identification algorithm for each frame is not an individual frame-based decision. Each occurrence of a detected person in each frame is considered a test sample which is classified using the knowledge about previously observed samples. This is because the concerned ratios are computed using the CC, FC, FR and FA scores, as described in Section 5.2, with the ground truths reduced to only persons and frames which were not missed:

$$CCR = \frac{CC}{CC + FC + FR}$$

$$FCR = \frac{FC}{CC + FC + FR}$$

$$FRR = \frac{FR}{CC + FC + FR}$$

$$FAR = \frac{FA}{FA + CR}$$

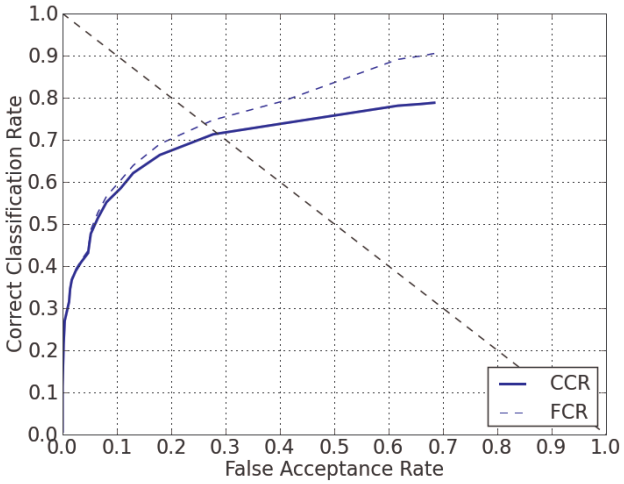


Figure 5.6: An example ROC plot for the CCR, FCR (and implicitly the FRR) in relation to the FAR in the case of identity tracking. In contrast to curves obtained, for example, in biometrics evaluations, the FAR may never reach the value 1, even if the acceptance threshold is set to 0.

with $CC + FC + FR$ and $FA + CR$ the total number of known and unknown persons, respectively, for which a corresponding hypothesis was output.

In summary, the consequences for the resulting ROC plots are as follows:

- Misses and false positives are not considered.
- The value of the FAR can stay well below 1, even if the acceptance threshold Th_{known} is set to 0. This is because identification features for some unknown persons present in the smart space may not be observable even once. Since persons are initially hypothesized as unknown, actual unknown persons in the space are always correctly rejected in such a case.
- The FRR also does not necessarily drop to 0, even if Th_{known} is set to 0, again because identification features for some of the focus persons may not be observable at all.

An example of a resulting ROC plot is shown in Fig. 5.6.

6 Experimental Evaluation

This chapter describes the database used to evaluate combined tracking and open set identification performance, and presents comparative results for the Joint Identity Tracking filter approach, under a variety of test conditions.

6.1 Evaluation Database

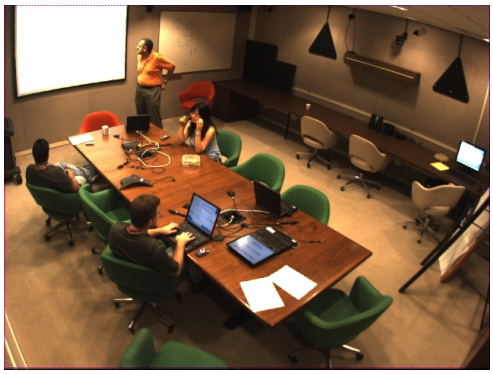
The Joint Identity Tracking method has been extensively evaluated on the Interactive Seminar database used in the CLEAR 2007 evaluations [103; 81]. This database features recordings of multiple users in realistic small meeting scenarios, captured in a variety of smart rooms equipped with a multitude of audio-visual sensors (see Figure 6.1). It offers five calibrated and synchronized visual streams, four from cameras mounted in the room corners and one panoramic ceiling-mounted camera, as well as synchronized audio streams from a minimum of four microphone arrays on the room walls. The dataset comprises 20 seminars from five recording rooms with varying audio-visual characteristics, with two annotated five minute segments per seminar, for a total of 200 minutes of recordings. In this dataset, a total of 67 individuals take part in small meetings, with typical meeting sizes of three to six persons. Of these 67 identities, 31 are learned in beforehand and constitute the set of known persons: 24 are trained in audio-visually, three are trained in using only the acoustic modality and four using only the visual modality. This is because the Interactive Seminar database only offers enrollment data for a limited amount of individuals. The remaining 36 "unknown" persons are those for which audio and visual data for enrollment or identification is insufficient. The ratio of known to unknown persons varies with each meeting, with a slightly greater number, on average, of unknown persons.

The visual annotations provide the 2D centroids in each camera view of the heads of each meeting participant, the bounding boxes of visible faces, the 3D head centroid locations obtained by triangulation, as well as a reference ID for each participant. These reference IDs stay constant throughout all recordings, such as a person appearing once in a seminar will have the same identifier assigned to it in the next. The audio annotations provide ground truth diarization information, including speech intervals for all active speakers, noise segments,

(a) ***AIT***

(b) ***UKA***

(c) ***ITC***



(d) ***IBM***

(e) ***UPC***

Figure 6.1: Scenes from the CLEAR 2007 Interactive Seminar database.

and segments of overlapping speech. Visual annotations are provided for each second of recording (every 15, 25, or 30 frames, depending on the framerate), and the segmentation accuracy of audio transcriptions is also approximately of the order of a second.

Although the recording setup was similar, the smart rooms figured in the Interactive Seminar dataset have quite differing characteristics, which tracking systems have to tune to automatically. For some sites, such ITC and IBM, colors are well distinguishable and offer a good cue for tracking. In others, such as UKA and UPC, they are quite hard to distinguish. In the UPC recordings, strong contrast to the smart room's white walls makes most colors appear as a nuance of black. In the UKA recordings, low overall illumination and strong chromaticity changes pose the biggest problem. In some rooms, the upper body detectors introduced in Section 3.2.2 function very accurately, delivering detections in at least one camera at almost frame rate for most participants. This is the case for the ITC room and especially for the UPC room (due to strong contrast with respect to the background). In others, upper torsos are only rarely detected, save for standing speakers. In the IBM room, for example, this is due to the fact that almost all cameras are placed at a high distance, making most room occupants appear quite small. For the UKA room, it is due in most recordings to the very low contrast to the background. In both cases, low resolution or weak illumination make visual identification of faces very difficult. Audio recording conditions also differ greatly. In the UPC room, very strong reverberations cause a problem for speaker localization and identification. In the case of the IBM room, the same is true due to a weak signal to noise ratio. A further important difference is sensor coverage. In the AIT recordings, for example, the observation space is constrained to a small area such that cameras are placed very close to occupants and provide a more narrow overview. As a consequence, most occupants are visible in at most two cameras, while for all other rooms, they are visible in four to five cameras most of the time. In the IBM room, the top perspective from the ceiling-mounted panoramic camera is heavily distorted, such that calibration information becomes inapplicable at the room edges. While most recordings were made using progressive scan firewire cameras, some rooms used analog cameras, which introduced interlacing artifacts into the views. Resolutions, focal lengths and framerates of the cameras, as well as the characteristics of recording microphones varied from site to site. The segments for evaluation were also taken from varying phases of the recorded seminars. While some recordings start with an empty room, with participants entering progressively, others start in the middle of the meeting, such that no controlled initialization in dedicated areas is possible. Trackers are expected to automatically detect and track multiple persons without any dedicated initialization phase, clean backgrounds or a-priori knowledge about person colors or attributes, for standing, sitting or walking persons alike. All these reasons

make the development of a tracker that functions robustly under all conditions without manual tuning or adaptation a quite challenging task.

The CLEAR 2007 Interactive Seminar database is used in the following experiments to evaluate the performance of the proposed open set identity tracking approach. The measures of interest that will be computed are the $MOTP$ and $MOTA$, measuring the overall tracking performance, and the $MITA$, measuring specifically identity tracking performance. The accuracy measures are computed as averages over all segments from all seminars of the Interactive Seminar database. Aside from providing meaningful statistics, this is also done to allow a better comparison to the official results of the evaluation workshops.

6.2 Experimental setup

The following experiments were conducted offline in a fully automatic, run-on fashion. Although some preprocessing steps were undertaken for the extraction of specific features, no batch processing was made, which means each tracking, classification or extraction module made only a single pass on the data and, for each time point, based its decision only on present and past observations. This is to provide that the results obtained in offline evaluation are still applicable to an online system.

The input data for the identity tracker consists for each recording of the visual streams from the room cameras as well as the audio streams from the wall-mounted microphone arrays. The metadata provided for each recording comprises:

- A set of background images for each camera view, taken before the start of the seminar, with no persons present.
- Sensor calibration information. For cameras, this consists of the extrinsic and intrinsic camera calibration parameters. For microphone arrays, it consists of their location, orientation and the internal configuration of their microphones (distance between pairs, etc). This information is necessary for projective transformation and source localization.
- The dimensions of the recording room. This is to provide reasonable boundaries for tracking.

Note that although information about sensor calibration or room dimensions could be used to deduce the recording site (and, e.g., tune detectors, thresholds, choose room specific classifiers or filter the results of identification), this was not done here. This also means that persons which only appear in recordings from one site can wrongfully be recognized in recordings from other sites, which makes the task more difficult.

For faster processing, all camera views are downsized to 320×240 pixels resolution (original resolutions include 640×480, 768×576 and 800×600, depending on the recording site and camera). In a live version of the identity tracker, implemented as a distributed system, this is also done to reduce network bandwidth requirements. The computation of foreground support maps, as described in Section 3.2.1 is made on an even lower resolution of 80×60, again for computational efficiency. This reduction in detail did not impact on the quality of tracking, and can be seen as a rough smoothing of already coarse and imprecise features. The cell size for the occupancy and occlusion grids described in Sections 4.1.4 and 4.1.5 is set to 10×10cm ($w_{cell} = 100$mm) and the exclusion and occlusion radii, R_{excl} R_{occl}, are set to 6 and 3 cells, respectively. These are intuitive values, considering an upper body width of 60cm, and have been empirically found to provide good results. The learnrate α for adaptation of identity models is set to 0.2 and the thresholds Th_{fg} and Th_{col} from Section 4.1.8 for creation (and deletion) of tracks are both set to 0.15, again based on empirical tests. Unless otherwise stated, for all experiments, the number of particles per track is set to 100. In several preprocessing steps, the localized upper torso detections, top view blob tracks and localized face ID features are extracted from the video sequences. Likewise, source localization is performed on the microphone array channels and combined with speaker identification results. Some of the associated details are described below.

6.2.1 Visual Recognition

Face recognition, as described in Section 3.3.2, is based on the approach presented in [36]. It has been evaluated under very similar conditions on the CLEAR 2007 Interactive Seminar database, which is also used here. The task was that of closed set identification for a subset of 28 individuals, using manually labeled face bounding boxes. In these evaluations, the task of aligning and identifying automatically detected frontal faces under the difficult conditions posed by the seminar recordings was judged too challenging and was not performed. Instead, separate training and test sequences of varying lengths with manually annotated face bounding boxes were provided for each target person. This means that the association of faces to persons was solved beforehand and the output of the identifier was the sequence level identity derived from all pre-segmented faces in a test sequence. Under these conditions, the recognition system achieved 84.6% accuracy for the hardest condition in terms of data availability (15s training segments, 1s test segments) and 96.4% for the easiest condition (30s train, 20s test). Those were, incidentally, the best results achieved in the CLEAR 2007 face identification task. These numbers should, of course, only be used as a rough orientation as to the accuracies that can be achieved on this dataset using state-of-the-art techniques.

For the evaluations done here, the task definition is by far more challenging, such that accuracies are expected to drop quite a bit: Here the same set of 28 individuals which was used for the closed set task is taken as the set of known identities and the trained classifiers from [36] are employed. However, the task is now that of open set identification, with 39 additional unknown faces appearing in the database. Moreover, the association of faces to persons for confidence accumulation is no longer known a-priori and has to be derived automatically. As for the above evaluations, the automatic detection and alignment of faces for recognition was not tackled. Instead, as before, the manually annotated face bounding boxes were used. Additionally, the annotated eye regions were employed to reject non-frontal faces. This is because the combined detection, alignment and identification of faces at such small resolutions is still an open research problem for which no satisfactory solution yet exists (see [103]). As described in [5], though, the problem can be circumvented e.g. by the use of active cameras that are steered to zoom in on target persons, capturing high resolution facial shots suitable for alignment and reliable identification. This could however not be investigated in the here presented offline evaluations, as no active camera views are available in the CLEAR 2007 database. As we will see later, even using manual annotations the recognition of faces proves very difficult. Another consequence is that face ID features, if available, come at a maximum rate of one face per second, the rate of the visual annotations.

As described in Section 3.3.2, the confidence scores for frame-based identification are derived from the distance scores of the k nearest neighbors of the test sample. To determine the optimal value of k and the resulting operating points for the open set classification case, a set of experimental runs was performed on the test dataset. These experiments evaluate strictly the identification performance, regardless of detection accuracy. For each detected face in all frames and camera views, an individual frame-level identification is made and the cumulative CCR, FCR, FRR and FAR scores are computed as described in Section 3.3. The resulting ROC curves are shown in Figs. 6.2 and 6.3.

As can be seen in Fig. 6.2, beyond $k = 10$ recognition performance decreases gradually with increasing values of k. This can most probably be explained by the non-linearity of the feature space in the classifier: Beyond a reasonable distance boundary, a large number of neighboring feature vectors from differing classes are included in the computation of the result. As the distance-based voting scheme is then no longer sufficient to reduce their influence, this decreases the quality of the computed confidence measure. From Fig. 6.3, we can see that decreasing the value of k below 10 brings no further improvement. On the contrary, for very small values of k, there are not enough neighbors for a meaningful application of the distance-based voting scheme, which makes the resulting confidence measure less reliable (for $k = 1$, the confidence measure is meaningless, as it is always 1).

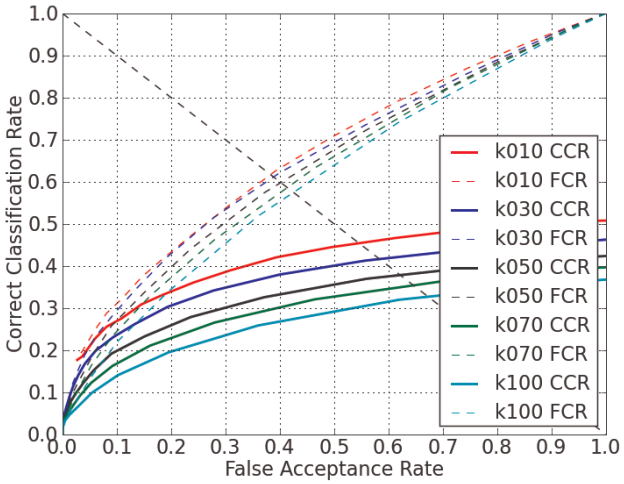


Figure 6.2: The ROC curves for the single frame open set face recognition performance, as a function of the number of samples used in k-nearest neighbor classification ($k = 10$ to 100). The best performance is achieved for a relatively small $k = 10$.

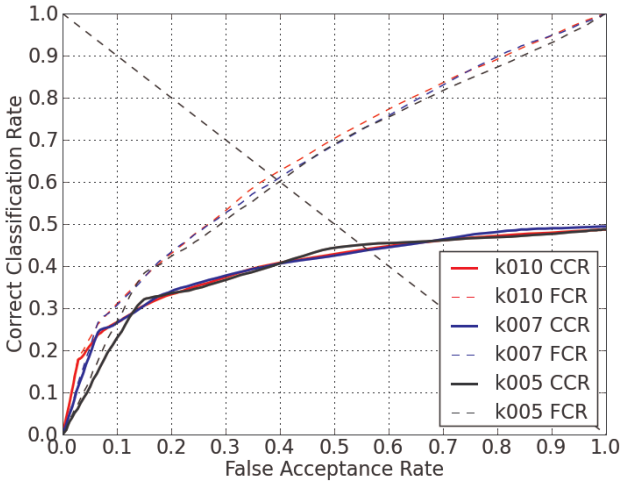


Figure 6.3: The ROC curves for the single frame open set face recognition performance, as a function of the number of samples used in k-nearest neighbor classification ($k = 5$ to 10). Decreasing the value of k below 10 does not yield any gain in performance.

Based on these results, for all following experiments, a value of $k = 10$ is chosen. Additionally, from the ROC curve at $k = 10$, the values γ_{max} and γ_{thresh} for warping of confidence values are determined (see Section 4.2.1). γ_{max} is taken as the threshold value for which the false acceptance rate drops to 0. The reason is that any test sample with a confidence value greater than γ_{max} must then be a positive test sample. As a consequence, γ_{max} represents the maximum value for which unknown person can still be falsely recognized. In this case, based on the plot, it is simply $\gamma_{max} = 1$ (This is not always the case, though). γ_{thresh} is chosen as the threshold for which at most 5% false acceptances are made. The reason for giving priority to low false acceptance rates (instead of choosing e.g. the rate at which $FAR = FRR$) is as follows: As identification is made on video sequences, many trials are performed per target person. Therefore, it makes sense to reject numerous identification attempts for various unknown persons, at the expense of rejecting a few trials also for a focus person. As results are accumulated, a few succesful trials are sufficient for a focus person to be correclty identified in the overall result. In this case, $\gamma_{thresh} = 0.8$.

6.2.2 Acoustic Recognition

The voice recognition algorithm, as described in Section 3.3.3 is based on the work of Jin et al., presented in [36]. As for the face recognition approach, it has been evaluated on the CLEAR 2007 Interactive Seminar database, in a closed set identification task involving 28 individuals and using pre-segmented intervals of clean speech. It reached 86.7% accuracy for the hardest testing condition (15s train, 1s test) and 99.1% for the easiest condition (30s train, 20s test).

Again, for the here presented evaluation, the speaker identification task becomes much more challenging:

- First of all, the segmentation of speech intervals from different speakers has to be made automatically. Suitable segments of clean speech must be detected and periods of silence ignored. In this approach, this is done by training an additional GMM for the "silence" class, which is used alongside the person-specific GMMs. Then, "segmentation by classification" is performed: The audio stream is continuously segmented into equal 1s intervals and identification is made on each. In this way, a fine-grained segmentation is achieved, silence periods are detected, and speakers are identified using one and the same technique. The short length of segments helps avoiding the accumulation of speech from alternating speakers into a same identification segment. It also, however, reduces the quality of results achievable on single segments, a fact which must be compensated later by accumulating observations belonging to the same speaker.

- Secondly, the association of speech segments to speakers has to be made automatically. This is done by calculating association probabilities for each present person, based on identity correlation and spatial proximity. For the latter component, the source of speech must be also be estimated automatically. Here, the system presented in [41] is employed. It was evaluated on the CLEAR 2007 Interactive Seminar database and reached an accuracy ($MOTA$) of 55% and a precision ($MOTP$) of 14cm. For the evaluations presented here, the source localization results from the original system are used without modification, such that the above numbers can be seem as directly reflecting the quality of acoustic localization available in this case.
- Thirdly, the identification task is posed as an open set problem: 27 speakers are trained in (of which 24 are also visually known, i.e. three known persons can only be identified using their voice). This means that 40 additional individuals occurring in the database are acoustically unidentifiable. As for the visual case, a confidence score is computed for each identification result by analyzing the n best GMM scores.

As for the visual case, experiments were conducted to determine the optimum size of n-best lists used in confidence estimation. The CCR, FCR, FRR and FAR scores were calculated for only identified audio segments which coincide with manually annotated segments of speech. This is again to avoid evaluating detection and segmentation quality. The results are shown in Fig. 6.4.

As can be seen, the identification performance rises with the value of n, up to the maximum number of 31 (the number of known speakers). This is in contrast to the face identification case, and is due to the different nature of the classifiers involved. In the following experiments, the value of n is therefore set to 31 and threshold values for confidence warping are chosen from the corresponding ROC curve. γ_{max} and γ_{thresh} are again taken as the threshold values for which $FAR = 0$ and $FAR = 0.05$, respectively. In the audio case, this results in the actual values $\gamma_{max} = 0.095$ and $\gamma_{thresh} = 0.075$. As can be seen, the normal range of confidence values here is much smaller, and rarely exceeds 0.1, such that the use of *unwarped* confidence values in modality fusion would be detrimental to the audio modality. As explained in Section 3.3.3, this is due to the nature of the underlying classification algorithm and depends also to a large part on the size of the n-best lists used in score normalization.

On a final note, one should consider that the acoustic identification is made here using just one audio channel from one of the wall-mounted microphone arrays. As training and test are made under matching conditions, acceptable results can still be achieved. The Person Identification tasks of the CLEAR evaluations have shown that using multiple microphones and combining classifier outputs can further boost performance on this dataset. Such a combination was, however, not attempted here.

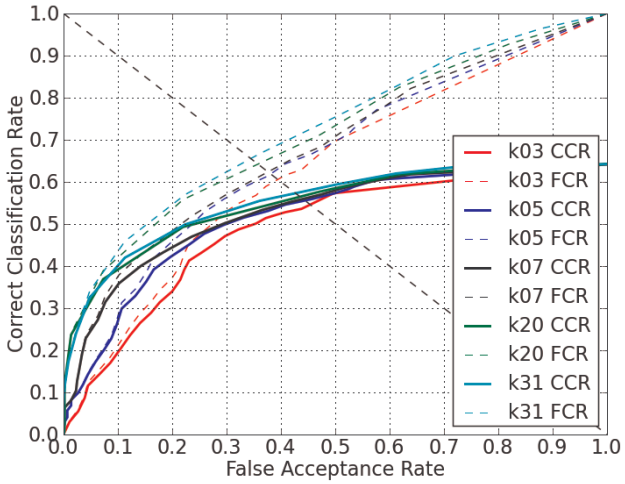


Figure 6.4: The ROC curves for open set speaker recognition on individual one second segments. The effect of the number of candidates included in the n-best list for score normalization is investigated. The best performance is achieved for the largest possible number of candidates $n = 31$.

6.3 Baseline

As a baseline to evaluate the advantages of the integrated ID tracking approach, a sequential algorithm was implemented which relies on an accurate detection and tracking step to estimate person identities. As explained in Section 2.3, the majority of approaches presented in the identity tracking field rely heavily on the flawless tracking of persons to maintain their identities in time. Identities are assigned directly to tracks, mostly at the start of the observation sequence, and are usually lost as soon as the corresponding tracks are terminated. Therefore, to offer a valid comparison, the baseline system is designed along the same principles.

The tracking algorithm for the baseline system functions just as for the *JIT* approach, using the same features, observation models, exclusion and occlusion principles, etc., but without its main extensions: No concept of "unobserved" particles is employed (save for the "exterior" particle, see Section 4.1.6), the belief in person states is not modified prior to resampling to account for "uncovered" observations, no "hidden persons" are modeled and identification is done individually for each tracked person.

The baseline system initializes a person model for each track and uses these person models as the basis for spatial association of ID cues. Here again, a person model comprises acoustic, visual and audio-visual pdfs for the modality-

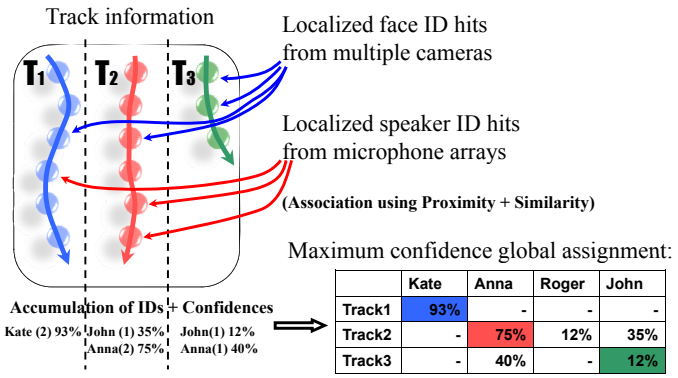


Figure 6.5: The process of mapping localized ID cues to person tracks in the baseline system. T_1 to T_3 represent the person tracker output. The blue and red arrows represent irregularly captured face and speaker ID cues, respectively. A spatio-temporal mapping is made for both types of cues, confidences are accumulated in time and a global assignment of IDs to tracks is made that optimizes the overall confidence level.

dependent modeling of identities. In contrast to the joint identity tracking approach, though, the location of an identity is not estimated probabilistically, but is directly given by the 3D coordinates of the corresponding track. The association of localized speaker and face ID features to person models is made based on spatial coverage and identity correlation. Features are mapped to the best correlating track. Features which cannot be associated to any of the available models (such as non-localized speech) are ignored.

After accumulation of observations and adaptation of ID models, the final identification hypothesis is not determined for each track independently, but rather by globally optimizing the hypothesis outputs for all tracks: For each model m, the identification confidence $P(id|m)$ is given by the corresponding audio-visual pdf. The problem of finding the assignment of distinct identities to person models that maximizes the overall identification confidence can be seen as a combinatorial problem (a maximum weight assignment problem), which is solved here using Munkres' algorithm [82]. The optimal assignment is recomputed every time a new identification feature is observed. The advantage of global assignment is that mapping of the same identity to several persons is avoided, as the system will change the hypothesized identity for one track based on new information for another track. This global post-processing of identification results is performed in the baseline system, as it represents a basic consistency

Data site	$MOTP$	$\overline{miss}$	$\overline{f.pos.}$	$\overline{mism.}$	$MOTA$
AIT data	20cm	14.51%	11.07%	0.43% (28)	74.00%
IBM data	19cm	14.40%	15.73%	0.27% (30)	69.60%
ITC data	14cm	8.61%	3.83%	0.15% (13)	87.41%
UKA data	18cm	24.10%	3.99%	0.45% (50)	71.45%
UPC data	13cm	5.67%	11.40%	0.59% (66)	82.34%
Total	17cm	13.59%	9.35%	0.39% (187)	76.67%

Table 6.1: Person tracking performance for the JIT approach on the CLEAR'07 dataset.

check which helps avoiding simple logical errors. It is by no means common, though, as the joint estimation of multiple identities has only seldom received attention in the literature (a recent exception is, e.g., [14]). Even so, designing a baseline system that performs no consistency check for multiple identities was not deemed appropriate for the comparative evaluations here. Figure 6.5 illustrates the process of data association and global ID assignment. The output of the baseline system, just as for the joint identity tracking system, are the hypothesized locations and identities of all persons in the space.

One obvious drawback of the baseline method is that only tracked persons can be identified and that learned identities are lost when the corresponding person tracks are lost. Identities and confidences then have to be reestimated as soon as tracking information is again available, from newly associated ID features. Additionally, in the baseline approach, ID features that cannot be associated to tracked persons are not considered, such that the information about the recognized identities is lost.

6.4 Evaluation Results

First, the tracking performance of the proposed JIT filter is evaluated. Table 6.1 shows the $MOTP$, $MOTA$ and associated component scores for individual sites (with 8 sequences per site), as well as the overall results.

As can be seen, the achieved accuracies are quite high, considering the difficult conditions of the CLEAR 2007 Interactive Seminar dataset. Tracking precision stays below 20cm in all cases. Tracking accuracy varies quite a bit for the different sites, owing to their specific difficulties, and a global $MOTA$ of 77.7% is reached. These results are comparable to the best tracking perfomance of the CLEAR 2007 evaluations, which were around 78.4% $MOTA$ with an $MOTP$ of $15cm$ (note that these are the numbers for the visual subtask in CLEAR, whereas here, acoustic features are also used for tracking. Note also that for

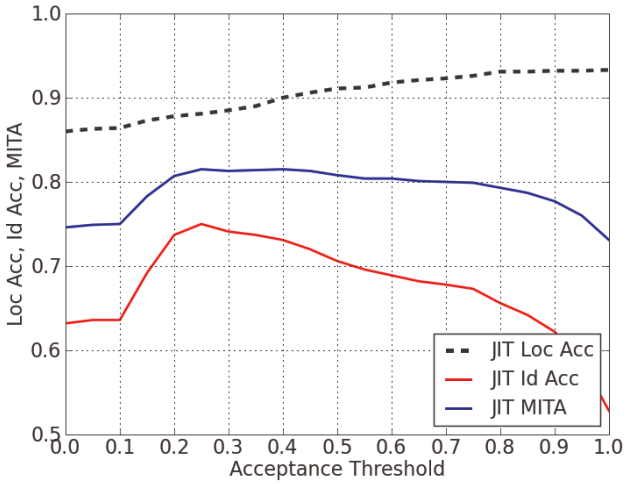


Figure 6.6: Multiple identity tracking performance of the JIT approach on the CLEAR'07 dataset.

the official evaluations, 2 seminars were not evaluated due to the lack of images for initialization of background models. These seminars are included here, with foreground segmentation initialized on-the-fly using the first 5 frames of recording. If these seminars are left out, the $MOTA$ rises slightly to reach 78.5%). The best perfomance is achieved for the ITC and UPC seminar rooms, which offer relatively good tracking conditions, with respect to colors, clutter, lighting, coverage and camera distances. The lowest performance is achieved for the IBM room, which was the only one where the top view from the ceiling camera was heavily distorted. This fact, and the relatively high distance of sensors to the observed seminar participants presented a big challenge to the tracker. What is also worthwhile to note is the relatively low mismatch rate, compared to the best official CLEAR results (0.79%, 361 absolute). This is due to a better management of track IDs, which avoids the creation of new tracks, if given observations can be associated to previously observed ones.

Next, the open set multiple identity tracking accuracy is evaluated in detail. Figure 6.6 shows the cumulative results for the LA, IA and $MITA$ for all seminars as a function of the acceptance threshold Th_{known}.

The highest reached MIT accuracy is 81.3% for a localization accuracy $LA = 88.5\%$ and an identification accuracy $IA = 74.1$. Roughly speaking, this means that 3 out of 4 known identities were correctly recognized, with no further false acceptance errors, etc, and that the location of these identities was almost always correctly found. These results are obtained for an acceptance threshold value of $Th_{known=0.3}$. Note that the LA and IA, taken individually, peak at different values of Th_{known}. The IA reaches and earlier peak of 75% for $Th_{known} = 0.25$, while the LA steadily rises, finally reaching a value of 93.3. The reason

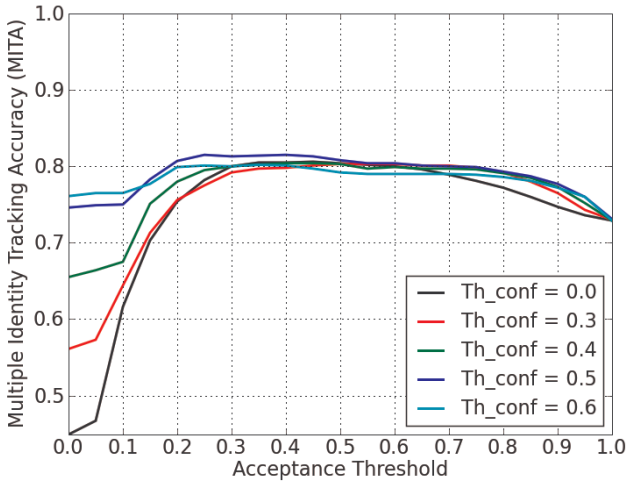


Figure 6.7: The $MITA$ obtained with thresholding of identification cues, based on their normalized, warped confidence scores.

for the steady rise of the localization accuracy is that, as Th_{known} increases, more and more focus persons are rejected as unknown, such that the chances of switching the locations of known identities decrease. Remember that no localization penalty is given if a focus person is correctly tracked, although not identified. Tracking *all* focus persons and rejecting all identities then results in an LA of 100%. This comes, of course, at the expense of a reduced IA, as seen in Figure 6.6.

6.4.1 Thresholding Identification Confidence

Section 4.2.1 explained how confidence values for identification are warped to allow a better fusion of modalities. Before identity tracking performance is evaluated in more detail, the use of confidence values is first investigated here with regard to temporal fusion. The idea is that since many low confidence identification results are generated by observations of unknown persons, a filtering of those may help avoiding the accumulation of errors, reduce the false acceptance rate, and therefore improve overall accuracy. This is done here by thresholding the warped face and speaker ID features. All features with confidence values below the threshold Th_{conf} are rejected entirely. Figure 6.7 shows the achieved MIT scores for different values of Th_{conf}.

As expected, when using no confidence thresholding at all, a high false acceptance rate causes a reduction in overall accuracy for lower values of Th_{known}. This effect is, of course, no longer observed when thresholding observations. What is interesting to notice is that the maximum achievable accuracy across

Data site	LA	IA	$MITA$
AIT data	86.0%	77.7%	81.8%
IBM data	88.5%	74.4%	81.5%
ITC data	90.7%	62.6%	76.6%
UKA data	79.5%	84.0%	81.8%
UPC data	91.2%	70.6%	80.9%
Total	88.5%	74.1%	81.3%

Table 6.2: Highest reached site-specific and global identity tracking scores for an acceptance threshold $Th_{known} = 0.3$.

all values of Th_{known} also increases due to the filtering ($Th_{conf} = 0.3$, 0.4 and 0.5). When setting the confidence threshold higher ($Th_{conf} = 0.6$), accuracies again start to drop. This is an indication that the feature warping, which attempts to normalize the distribution of confidence values to the range $[0, 1]$, with the value of 0.5 as acceptance threshold, is indeed effective. On average, identification features with warped confidence values below 0.5 can be considered as coming from unknown persons, and can therefore be rejected. For this reason, the accuracy curves in Fig. 6.6 as well as all other results presented in the following have been obtained using confidence thresholding with $Th_{conf} = 0.5$.

6.4.2 Identity Tracking Accuracy

As shown in Fig. 6.6, the maximum value for the $MITA$ is reached around $Th_{known} = 0.3$. For this value, the individual site-based and overall MIT scores are computed and presented in Table 6.2.

A few things can be seen from the comparative table. First, the highest localization accuracies, at around 91% are reached for the ITC and UPC recordings, which also provided for the highest $MOTA$ scores. The values are even significantly higher, due to the fact that for the LA, only focus persons are considered in the computation of the metric. The values for other sites are also quite high, though (on average, 88.5%), and do not necessarily correlate with their individual $MOTA$ scores (this is most obvious in the IBM case), for the same reasons. In contrast to localization, the highest identification accuracy is reached on UKA recordings. This is because in most of these recordings, of the seminar participants present, only one or two are focus persons, and these usually occupy the function of the main presenter, such that they can be easily identified acoustically. In contrast, the lowest IA score is reached for the ITC recordings, where the number of focus persons often exceeds that of unknown ones (for 4 of the 8 recordings, the number of unkown persons is even 0). This makes it easier to make identification errors, especially in the case some focus persons rarely

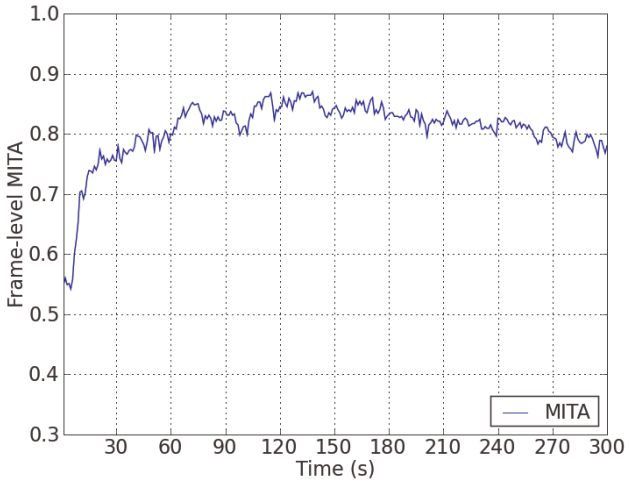


Figure 6.8: The evolution of the $MITA$ as a function of time.

speak or are difficult to observe. Except for the ITC site, all $MITA$ scores are around 81%.

Next, the evolution of frame-level MIT-scores is shown in Figure 6.8, using the acceptance threshold $Th_{known} = 0.3$. The horizontal axis of the plot represents the time axis, going from second 0 to second 300. The value for each second is the average MIT-score for this second over all sites and seminars. This representation is possible since all evaluation segments have a length of five minutes. As can be seen, the identity tracker needs a few seconds to initialize tracks and observe identities. After 60 seconds, a MIT score of 80% is reached, and the accuracy still rises slowly in the following minutes, reaching at times 87%.

In the last two minutes, one can observe a slight decrease. This is due to the fact that, as time progresses and noisy observations come in, the chances of falsely identifying unknown persons increase, causing the FAR to rise. Although this is accompanied by a parallel increase in the CCR and decrease in the FRR, the effect of the rising FAR cannot be fully compensated. Throughout the 300 second time window, the average false classification FCR, false positive rate FPR and, from second 50, also the miss rate MSR stay nearly zero. Figure 6.9 shows the evolution of the relevant non-zero scores on the time axis, assuming $Th_{known} = 0.3$.

Finally, Figure 6.10 shows the ROC curves for correct classification and false classification averaged over all sequences. These curves are calculated using the results of identity tracking, as described in Section 5.3.

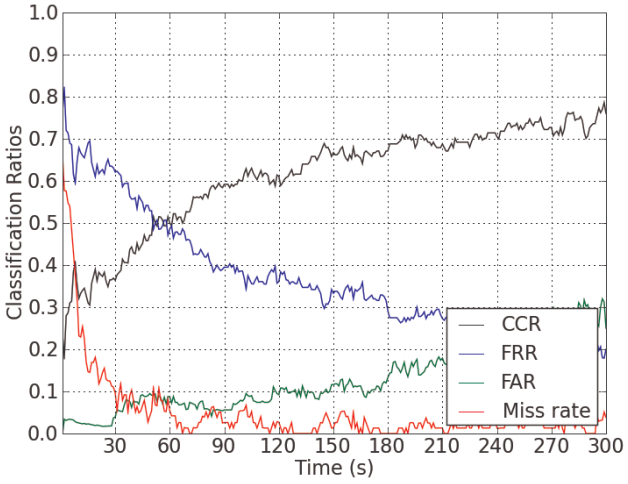


Figure 6.9: The evolution of component ratios as a function of time. The FCR and FPR are close to zero and are not plotted.

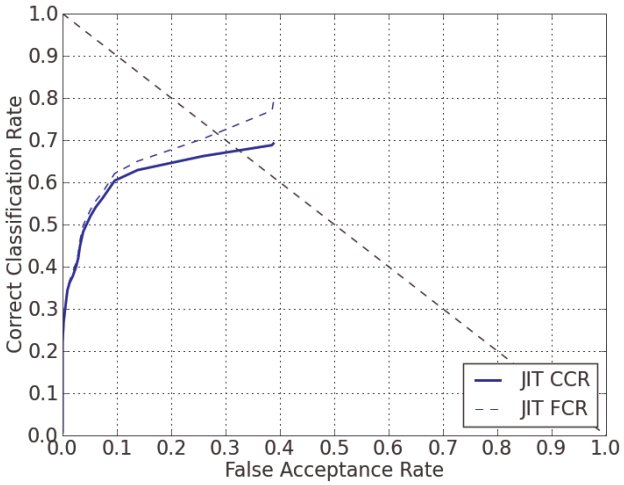


Figure 6.10: The ROC curves for the joint identity tracking task on the CLEAR'07 seminar data.

As explained in Section 5.3, the false acceptance rate does not exceed a value of 40%, even when setting Th_{known} to 0. The equal error rate for the CCR is reached at around 32%. This is not, however, the operating point chosen for the computation of absolute scores presented above. For a value of $Th_{known} = 0.3$, the actual scores are $CCR = 56.3\%$, $FRR = 42.3\%$ and $FAR = 7.4\%$. This is in line with the projected value of 5% FAR, used for the selection of parameters in the feature warping step (see Section 6.2). Compared to the frame level visual and acoustic ROC curves presented in Sections 6.2.1 and 6.2.2, the curves shown here rise much faster, with the CCR reaching 60% already for an FAR of 10% (compared to around 27% and 40%, for single frame visual and acoustic identification, respectively). This shows the advantages of temporal and modality fusion based on the continuous tracking of identities.

6.4.3 Baseline Comparison

In this section, the proposed JIT filter framework is now compared to the baseline system described in Section 6.3. Figures 6.11, 6.12 and 6.13 show a comparison of identification accuracy, localization accuracy, and identity tracking accuracy, respectively.

In Fig. 6.11, the first comparative curve ($Best$) shows the case of undisturbed tracking results. As could be expected, the two approaches perform equally well over almost the entire range of values of Th_{known}, with a very slight advantage for the JIT approach. Curves for the localization accuracy are also very similar. The reason is that when tracking results are reliable, such that all focus persons are localized at all points in time, performing identification based directly on those results, i.e. only for tracked persons, is a viable option. The true strength of the JIT filter approach, in comparison, lies in its robustness to detection and tracking errors.

The second comparative curve ($ErrFG$) is shown for the case the foreground feature is heavily perturbed, such that it is unsuited for tracking. This can happen for a variety of reasons, in realistic scenarios, including e.g. cases where lights are dimmed during a slide show presentation, where doors, or window blinds are opened or closed, where many non-human objects are frequently moved, etc. As the recordings themselves can no longer be changed, it is realized here artificially by setting the adaptation rate for the background model to an unreasonably high value. As the foreground feature is an important component for deciding when to keep tracks alive, the perturbation of the foreground feature causes a heavy degradation of tracking performance.

The third comparative curve ($NoCams$) shows the extreme case where all cameras are unavailable, such that visual tracking is wholly impossible (Of course, this means that the visual modality is unavailable also for identification). Note

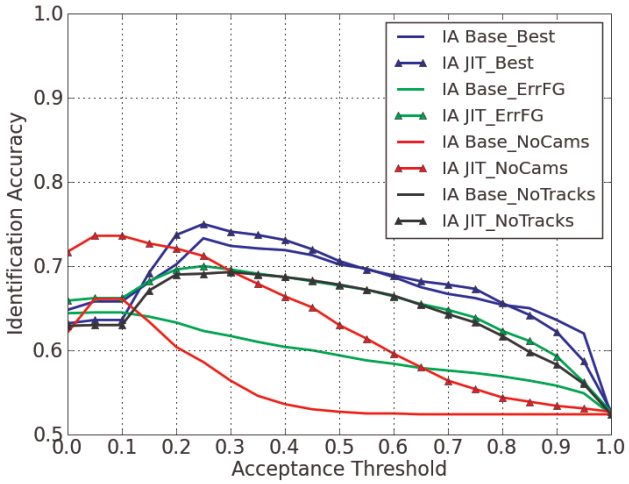


Figure 6.11: Identification accuracies for the *JIT* and the baseline approach in the presence of component failure. The identification accuracy for the baseline system in the case no tracking is possible (*Base_NoTracks*) is always zero, which is why the corresponding curve is not visible here.

that tracking of multiple persons is still possible using the acoustic modality alone, by building on the results of 3D sound source localization.

The fourth comparative curve (*NoTracks*) shows another extreme case where tracking of person locations can not be realized, both visually or acoustically: In many natural scenarios, calibration information is simply not available for the observing cameras and microphones, either because a calibration at global scale is infeasible, or because the configuration of sensors changes too rapidly. In such a case, the estimation of 3D tracks, based on 2D image features or microphone array azimuths is not possible. Here, the extreme case is simulated by preventing the formation of tracks in the first place, such that ID features must be accumulated and identities estimated without the spatial correlation component.

As can be seen in Figure 6.11, as the quality of the foreground support drops (*ErrFG*), the difference between the approaches becomes clear. The *JIT* filter keeps identities and confidences even through tracking errors, while the baseline system loses information every time a track is lost and has to be reacquired. This is partly due to the modeling of "hidden" persons, which are still hypothesized present although not observed.

When no visual information is available (*NoCams*), the modeling of "unobserved" tracks, which are not propagated, allows the *JIT* approach to maintain the positions of previously observed person, such that a tracking of multiple

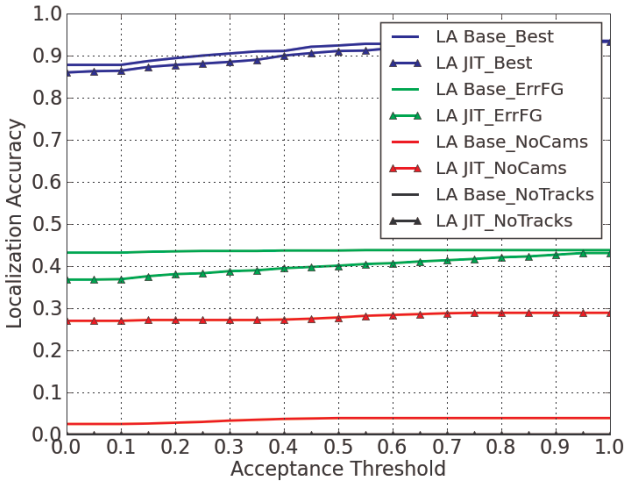


Figure 6.12: Localization accuracies for the JIT and the baseline approach in the presence of component failure. The localization accuracy for both systems in the case no tracking is possible ($NoTracks$) is always exactly zero.

alternating speakers is still possible to some extent. This is not possible for the baseline system, which loses track of individual speaker locations shortly after each observation. While the main effect is a significant decrease in localization accuracy, as shown in Fig. 6.12, it also affects identification accuracy, as identities for focus persons have to be held based only on their speaking frequency. In the case no cameras are used, The IA curves in both cases reach their maximum for a lower value of Th_{known} of around 0.1. This is because the acoustic modality causes fewer false acceptances, measured in numbers of persons, than the acoustic one, as it is often observed only for a few main speakers. This means the acceptance threshold can be set much lower without increasing the FAR significantly. The highest reachable identification accuracy stays below that for the multimodal case, though.

Finally, When no tracks can be generated at all, the JIT filter achieves comparable identification accuracy as in the presence of heavily flawed foreground support. The highest reached accuracy is 70.0%, for $Th_{known} = 0.25$, although no spatial information can be used for data association, for multiple persons in an open set identification scenario. For the baseline system, as no track information is available, the IA also drops to 0.

In Figure 6.12, the localization performance can be seen in more detail. Both systems achieve a high accuracy in the absence of tracking perturbations ($Best$), with slightly better results for the baseline system. This is in opposition to the IA scores in the same case, and again comes from the fact that the LA and IA are in opposition in the case of correctly recognized but badly localized

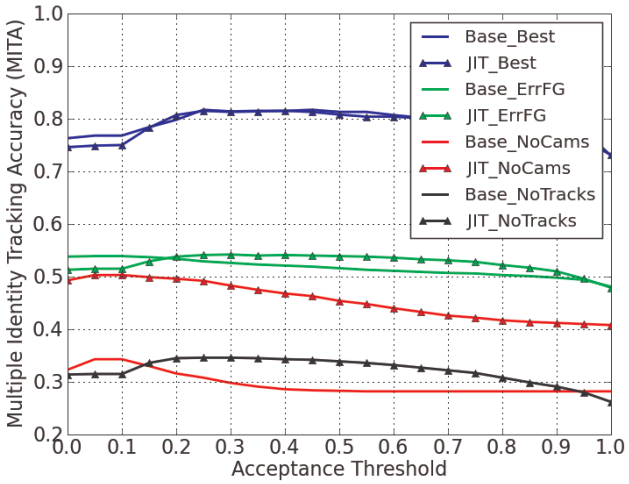


Figure 6.13: The $MITA$ scores for the JIT approach and the baseline system as feature extraction, sensors, or whole modalities fail. Accuracies for the JIT approach degrade gracefully in the face of component failure. The MIT accuracy for the baseline system in the case no tracking is possible (*Base_NoTracks*) is zero for all values of the acceptance threshold, such that the corresponding curve is not visible here.

identities: If the identity is rejected, it results in a decrease of the IA (through an increase of the FRR), but also in an increase of the LA, in the case the focus person is still correctly tracked. Both types of errors come to a balance in the $MITA$, as can be seen in Fig. 6.13. The effect is even stronger in the case of erroneous foreground support ($ErrFG$), although the relative increase in the IA (for the JIT approach compared to the baseline) is stronger than the relative decrease in LA. When the visual modality fails completely ($NoCams$), the localization accuracy for the baseline system drops considerably while, as explained above, the JIT approach is able to keep basic estimates for at least a subset of focus persons. Finally, when no 3D tracks can be generated, the LA drops to 0 in both cases.

Figure 6.13 sums up the results for the case of degrading tracking performance. While both approaches initially show comparable performance, the JIT filter approach already shows a slight advantage in the case of flawed foreground support. The difference becomes much larger as the video modality fails, and is largest as tracking information becomes completely unavailable. The advantage of the JIT approach is that although information about locations becomes increasingly inaccurate, it is still capable of preserving the information about identities, while in the baseline approach, both are tightly coupled, such that accuracies tend toward 0 as tracking degrades.

Modality	$MOTP$	$\overline{miss}$	$\overline{f.pos.}$	$\overline{mism.}$	$MOTA$
Audio only	23cm	84.15%	12.08%	0.47% (228)	3.30%
Video only	17cm	13.63%	9.81%	0.40% (195)	76.16%
Audiovisual	17cm	13.59%	9.35%	0.39% (187)	76.67%

Table 6.3: Person tracking performance for the purely acoustic, the purely visual, and for the audiovisual cases.

6.4.4 Modality Fusion

Here, the effects of audio-visual modality fusion, which could already be perceived in the previous section, are evaluated in detail. Table 6.3 shows the MOT accuracies reached for the single modalities as well as the audio-visual fusion.

The MOT results for the acoustic only case are obtained by disabling all cameras completely. The results for the visual only case are obtained by disregarding all acoustic features. As can be seen, visual tracking clearly outperforms acoustic tracking, simply because of the constant availability of observations simultaneously for all persons. In the acoustic only case, temporary tracks for active speakers can still be maintained, though, such that accuracies do not drop to zero. The addition of the acoustic observations to visual tracking still allow a slight gain in performance.

Figure 6.14 now shows the MIT accuracies reached for audio, visual, or audio-visual *identification*. In contrast to the above evaluation, the visual *tracking* features are used here in all cases (the cameras are not disabled completely). The difference lies solely in the inclusion or not of speaker ID or face ID features. Multimodal fusion is made here using the product rule.

The audio modality clearly outperforms the visual one, reaching a maximum MIT accuracy of 79.7% for $Th_{known} = 0.2$. The result is not surprising, considering the frame level accuracies reached by both systems (see Sections 6.2.1 and 6.2.2). This is also in line with the observations made in the official 2007 CLEAR evaluations, concerning the Interactive Seminar dataset [103]. Although the visual modality offers the advantage of continuously observing all smart space occupants simultaneously using several overlapping sensors (as opposed to the acoustic case, where only one person can be identified at one time), this is outweighed by the difficulties posed by extremely low face resolutions. Even using manual annotations for face detection and alignment, the potential of face identification can not be fully exploited and performance stays below that of the acoustic case. The lower IA scores are compensated to some extent, though, by a higher localization accuracy, such that a maximum MIT score of 76.0% is reached for an acceptance threshold $Th_{known} = 0.25$. The fusion of both

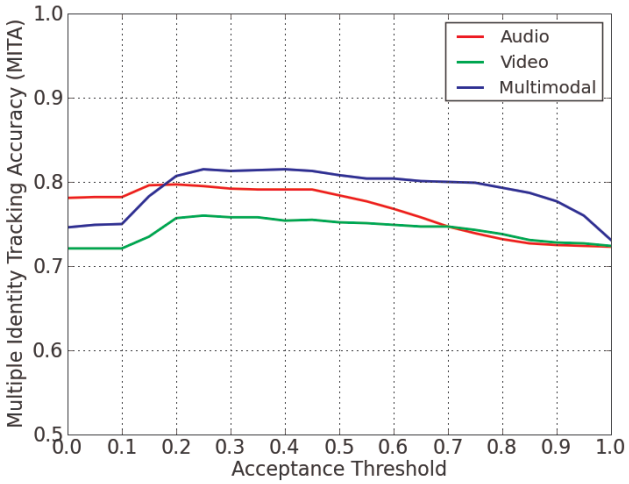


Figure 6.14: $MITA$ scores for single modality acoustic or visual and for multimodal identification.

modalities, finally, offers a clear advantage. The multimodal system achieves a maximum MIT score of 81.3% for $Th_{known} = 0.3$.

The advantages of multimodal fusion become even better apparent when considering the correct and false classification ratios. Figure 6.15 shows the corresponding ROC plots. Using both modalities, clearly higher CCR rates can be achieved than by using any one modality alone. Although the curve for the audio only case rises very sharply for low values of Th_{known}, it is soon caught up with by the multimodal case, with both roughly achieving 60% CCR for a false acceptance rate of 10%. The additional advantage of the audio-visual system is that its operating point can also be better adjusted, with an possible FAR of up to 40%. In the audio case, correct classification rates (as well as false classification rates) are limited, even when setting $Th_{known} = 0$, due to the fact that persons may never be identified (correctly or falsely) even once, based on their voice. This is compensated in part, by the visual modality, although additional errors are also introduced. The audio-visual system achieves an equal error rate of $CCR = FAR = 32\%$.

The drawback of performing multimodal fusion, which needs to be carefully balanced, is that false acceptance errors (and false positives) from both modalities are accumulated. Unknown persons that never spoke, e.g., and therefore were never falsely accepted based on the acoustic modality, are now recognized based on their faces, and vice versa. This is the reason the difference in accuracies is not more clearly apparent in the overall $MITA$ score, which weighs all errors equally. One must remember, however, that this result is dependent on the evaluation dataset. For other scenarios, where e.g. the percentage of persons that are identifiable using both of the two modalities is lower, the complementary

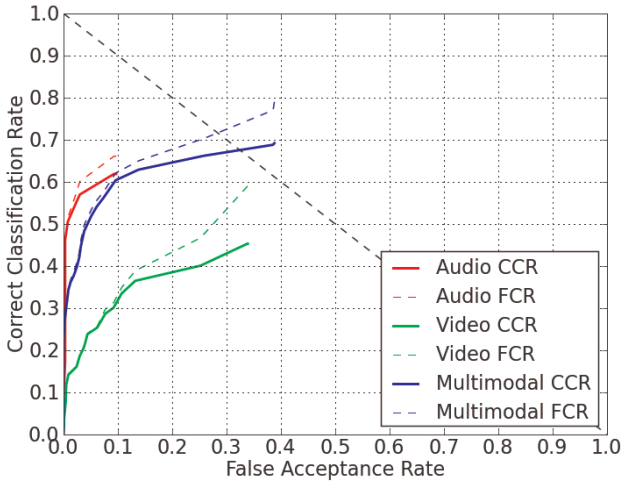


Figure 6.15: ROC curves for single modality and for multimodal identity tracking.

nature of face and voice identification may be even more apparent. Figure. 6.16 shows the evolution of the $MITA$ for the audio, the visual and the multimodal case, as a function of time. This representation is possible as all sequences in the dataset have a length of 5 minutes. As explained in Section 6.4.2, the slight decrease in accuracies, on average, in the second half of the recordings is mainly due to the fact that, as more and more observations become available, the chances of observing the faces or voices of unknown persons and wrongfully identifying them increase. This is reflected by a slight increase of the false acceptance rate and consequently a decrease of the MITA.

Although this was not the case for the Interactive Seminar dataset, the fusion of modalities can, in principle, lead to a faster overall identification result for all occupants. As faces are not observable, voices may be identified, and vice versa. This may be a great advantage in scenarios with lower sensor coverage, where certain key identities need to be localized quickly and one cannot wait, e.g., for a specific participant to face a camera (or to take his turn speaking) to make an identification. This aspect could be investigated further in the future, using more balanced datasets.

Next, we will now evaluate the effects of different fusion strategies for multimodal identification. Figures 6.17 and 6.18 show the $MITA$ curves and the CCR - FCR plots, respectively, for 3 different fusion strategies:

1. Using a single audio-visual model to accumulate both speaker and face ID cues.
2. Using separate models and combining their results using the weighted sum rule. Here, equal weights of 0.5 are assigned to each modality.

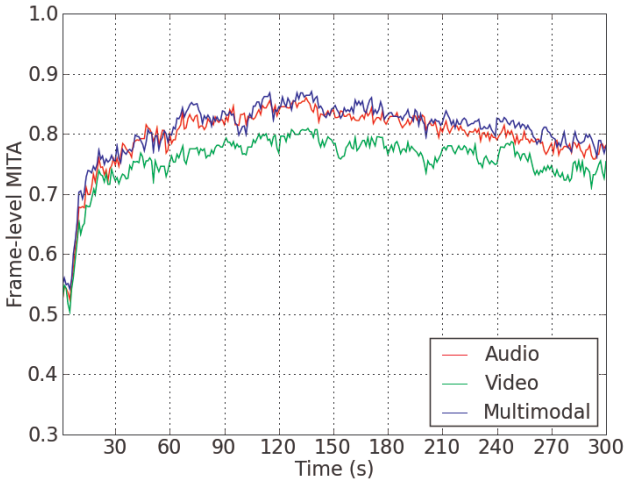


Figure 6.16: Evolution of the frame-based Multiple Identity Tracking Accuracy (MITA) in time, averaged over all seminars and segments, for audio, visual, and audio-visual identification.

3. Using separate models and combining with the product rule.

As can be seen, the sum and the product rule outperform the single-model strategy for a broader range of acceptance thresholds. As explained in Section 4.2, this can be explained by the fact that when using a single model, results may be biased toward one modality based on observation frequency, such that the advantages of fusion are not fully exploited. When using separate models, the product and the sum rule achieve comparable results. The advantage of the product rule is that it is the least sensitive to the acceptance threshold Th_{known}. This is of course an important point in realtime applications where it is not possible to determine the optimal threshold for a yet unseen scenario beforehand.

6.4.5 Temporal Fusion

This section analyzes the effects of temporally fusing identification results on identity tracking performance. For this, the *JIT* framework is modified such that the identity of a person (tracked or hidden) is determined based only on the last associated ID observation (*NoTempFusion*). All previous observations for the same person are disregarded. This is realized by setting the learnrate α for identification to 1. Separate models for the audio and visual modalities are still kept, though, and the fusion made using the product rule. Figure 6.19 shows the resulting *LA*, *IA* and *MITA* curves, as well as the open set ROC curves for the standard *JIT* approach (*TempFusion*), and the modified approach (*NoTempFusion*).

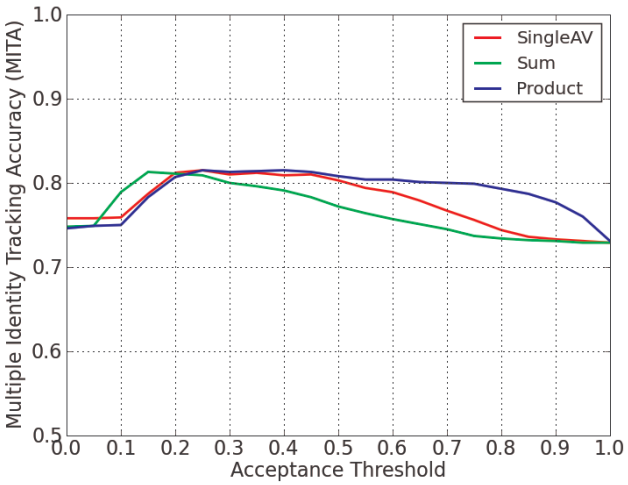


Figure 6.17: A comparison of MIT scores reached in the multimodal case, when using the sum rule, the product rule, or without separation of identity models.

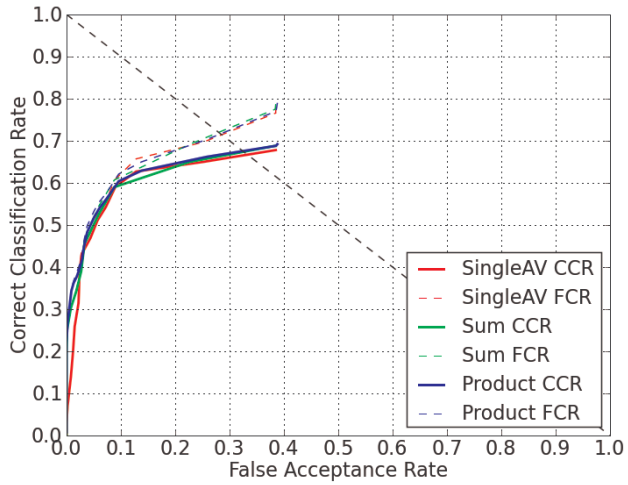


Figure 6.18: The ROC plots for comparison of the single model strategy, the sum rule and the product rule for multimodal identity tracking. No significant difference can be observed.

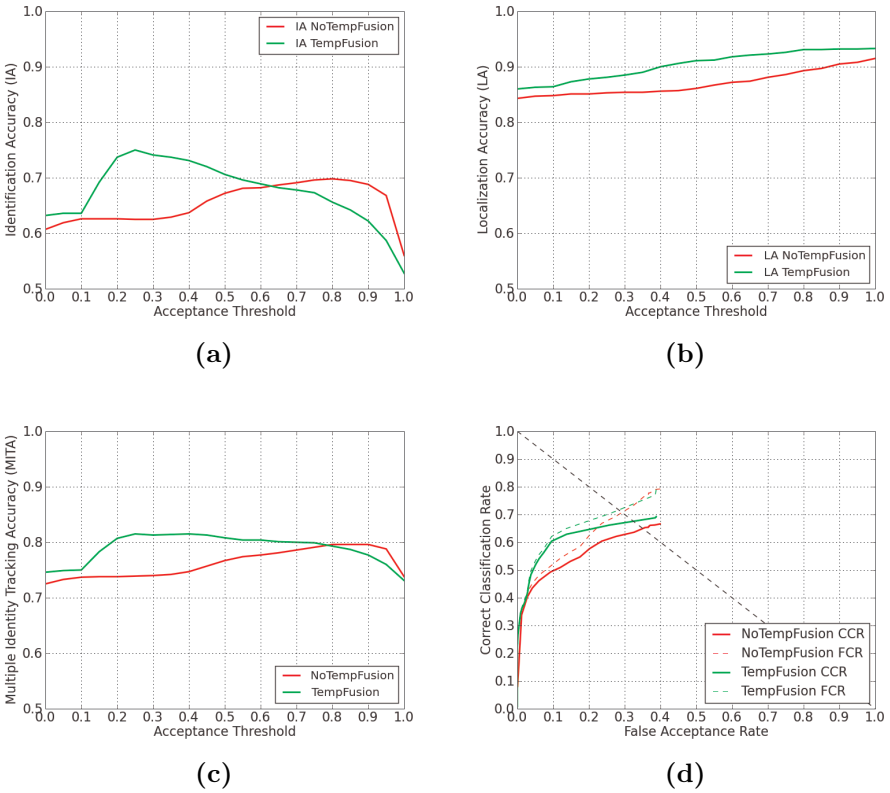


Figure 6.19: The effects of confidence based temporal fusion on identity tracking accuracies. Both the MIT scores and the open set ROC performance are significantly increased when temporal fusion is made.

Failure	$MOTA$	LA	IA	$MITA$
None	76.67%	88.5%	74.1%	81.3%
Color	63.28%	83.5%	73.6%	78.6%
Det+Top	43.22%	60.7%	72.3%	66.5%
Foreground	25.68%	38.8%	69.6%	54.2%
Cam5	60.49%	71.6%	71.8%	71.7%
Only Cam1&2	26.90%	54.1%	74.0%	64.1%
No Cams	3.30%	27.2%	69.4%	48.3%

Table 6.4: Person tracking and Identity Tracking performance in the presence of individual component failure for the multimodal case.

According to expectations, both the localization and identification accuracies are clearly higher when performing temporal fusion. From Figs. 6.19(a) and 6.19(c), one can see that the operating point for open set identification (the acceptance threshold Th_{known}) has to be set extremely high to achieve the best performance, when no temporal fusion is made. This is because false acceptance errors for unknown persons can only be rejected based on the confidence values for single identification results. They can no longer be averaged out using the information about possibly conflicting results in an observation sequence. The ROC curve shows that temporal fusion also yields a superior system with respect to correct classification rates.

6.4.6 Graceful Degradation

In this section, the performance of the JIT approach is evaluated in the presence of various types of failures. These include failures in the feature extraction process, as well as complete failures of observing sensors. For this, several test runs were performed, by disabling the color feature ($Color$), by disabling upper torso detection and top view blob tracking ($Det + Top$), by perturbing the foreground support maps as in Section 6.4.3 ($Foreground$), by disabling the top view (and all associated features) completely ($Cam5$), by using only two cameras ($Cam1\&2$), and by disabling all cameras completely ($NoCams$). The results are shown in Table 6.4.

As can be seen, the failure of individual feature extraction components causes a sensible reduction in the MOT performance. By disabling colors and detections it drops from 77% to 63% and 43%, respectively. The most important feature for tracking is still the foreground support. If it fails, accuracies drop further to around 26%. It should be noticed that the drop in the MIT scores is not proportional, though. This indicates that as tracking quality decreases, the tracker is still capable of maintaining relevant information about focus persons,

Particles	$MOTA$	LA	IA	$MITA$
300	75.78%	89.1%	74.0%	81.6%
100	76.67%	88.5%	74.1%	81.3%
50	73.95%	86.9%	73.7%	80.3%
25	67.61%	77.5%	73.3%	75.4%

Table 6.5: Person tracking and Identity Tracking performance with varying numbers of particles per track.

while information about unknown persons is more quickly lost. When disabling the top view, as well as when using only two cameras, cams 1 and 2, MOT accuracies also drop considerably, while MIT scores are less affected. The reason for evaluating against a loss of the top view is that it offers the best features for person tracking, showing a mostly unoccluded view of the whole scene. The evaluation using only two fixed cameras serves only for demostration purposes. Any other two cameras could have been chosen, or an average of all possible combinations taken. This was not deemed necessary here, as the effects can be well exemplified by the investigated case, and should not differ greatly in the others. In any case, a lower bound is given by the case where all camera views are unavailable ($NoCams$). There, a MIT score of 48.3% can still be achieved. While the LA score gradually decreases with component failure, the IA score stays relatively unaffected, rarely dropping below 70%. This is one of the main advantages of the JIT filter framework, which is flexible to feature extraction, sensor and modality failure. It is capable of providing relevant information concerning identities in the smart space, as long as some of the underlying tasks of person tracking, source localization, face recognition, or speaker identification, can be accomplished with a sufficient degree of accuracy.

Finally, the effect of the particle cloud size on identity tracking performance is tested. The results are shown in Table 6.5.

The first column in the table stands for the number of particles per track. The results show that increasing the number of particles to 300 does not yield any significant improvement. The $MOTA$ values even drop very slightly, due to a small increase in the false positive rate. Reducing the particle mass by half also causes only a very slight decrease in accuracies. Even using only 25 particles per track, scores of 68% $MOTA$ and 75% $MITA$ are reached. Since the number of particles directly influences the runtime performance of the tracker, being able to achieve high accuracies with fewer particles can be essential for its deployment in realtime applications, involving high numbers of users.

6.5 Discussion

This chapter presented an evaluation of multiple user open set identity tracking performance on the CLEAR 2007 Interactive Seminar database. The results show that a multimodal analysis is indeed possible for several simultaneous smart space occupants using only distantly placed sensors, and that it poses several advantages compared to unimodal analysis. This is with respect to the overall accuracies reached, as reflected in MIT accuracies and correct classification rates, but also with respect to the flexibility offered when single modalities are not usable.

The experiments also show that acoustic identification can be a very powerful source of information, especially when used in conjunction with microphone array source localization. High identity tracking accuracies can be achieved for multiple persons using only the acoustic modality for both tracking and identification. Even so, the conditions for the acoustic case must still be considered relatively controlled, as only indoor scenarios with relatively few sources of noise are evaluated. The potential of the video modality, on the other hand, can still not be completely exploited, as low resolutions, lighting changes, and limited sensor coverage still pose severe problems for the detection and identification of faces. One obvious solution is the use of active cameras that are steered toward targets of interest and acquire high quality, high resolution views of faces. This solution was investigated with success in [5] and is also applicable to the JIT framework proposed here.

It has also been shown that the proposed JIT approach degrades gracefully with individual modality failure, tracking failures, reduced sensor availability, and limited amounts of particles. Especially in those cases, it outperforms the baseline approach, which performs standard SIR particle filtering and builds on the results of tracking for identification. Even as localization becomes increasingly difficult, the JIT approach maintains important information about focus persons, with identification accuracies staying above 70% in almost all cases.

6.6 Live System

This section presents a realtime implementation of the JIT approach. It is realized as a live system, with individual processing components distributed over a network of computers. The system utilizes the input streams from four room corner cameras, one ceiling-mounted top-view camera, four T-shaped microphone arrays on the room walls, one omnidirectional microphone on a central meeting table, as well as two steerable SONY EVI-D70P cameras.

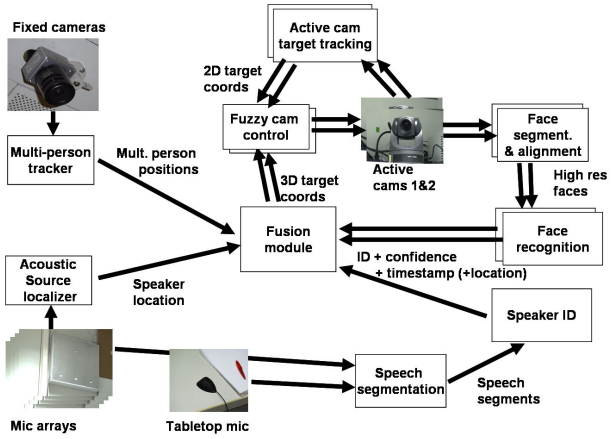


Figure 6.20: Overview of the components of the live multimodal identity tracking system.

All components and sensors work together to achieve the goals of high room coverage, precise localization and quick and accurate identification. Each component is designed to be fully automatic and realtime-capable, and is seamlessly integrated in the overall system. The prime building block is the *JIT* filter estimating the locations and identities of all persons present in the room. Camera observation modules preprocess the image information locally and extract upper body detection and foreground support maps, which are sent to the fusion module. Modules for speech detection, segmentation and speaker identification, coupled with a source localizer using the microphone arrays inputs, deliver precisely localized ID cues whenever a speaker becomes active. The pan-tilt-zoom cameras focus in on tracked persons to gain high resolution snapshots usable for face identification. The faces of target persons are automatically detected and aligned in the active camera images. This information, together with continuously updated camera calibration parameters, is used for 3D localization of the face. The central *JIT* module analyzes the output of all components and implements a variety of strategies for target and camera selection. It performs spatio-temporal association of incoming identification cues to person models, accumulates statistics in time, and continually optimizes the global scene configuration according to the sequence of observations.

A total of eight Pentium IV, 3GHz machines is used: Five for the visual tracking, one for the acoustic tracking and identification, and two more for the control of active cameras and the tracking and identification of faces in closeup views.

Fig. 6.20 gives an overview of the system and of the interaction between its components.

6.6.1 Active Camera Face Capture

As opposed to the offline implementation presented in previous chapters, the live system makes use of active cameras to better exploit the potential of face identification. Based on the current room occupant configuration, and the confidence in each person's identity, the fusion module decides on the actual persons of interest and on the active cameras to be used to acquire frontal face views for identification. The target persons' 3D scene positions are sent to the active camera tracking and control modules where several subtasks are accomplished:

- The control of the camera's pan, tilt, and zoom factors to focus in on the desired region.
- The detection and alignment of frontal faces whenever available and the reprojection of detection hits to the 3D scene.

The face acquisition task is accomplished by two SONY EVI-D70P cameras mounted on the room walls. They are placed such as to offer good views of a person giving a presentation in front of the projection board or coming in the door, but also offer good coverage of the audience and the rest of the room. Each camera is connected to a separate machine running dedicated components for control, detection, alignment and identification of faces.

Face Alignment and 3D Scene Reprojection

Once an active camera has zoomed in on a target person, high resolution snapshots of frontal views of his or her face are taken, aligned for face identification, and their 3D scene position is estimated. The detection of frontal faces is made using the cascaded Haar-feature classifiers proposed by Viola et al. in [113]. The inside of the detected face rectangle is scanned in a second pass with Haar-feature classifier cascades trained to detect eye regions. Only if two eyes can be detected, reasonably situated inside the face rectangle, is the aligned face passed on for recognition. Although this may cause some faces to be discarded because both eyes could not be detected, the two stage approach does guarantees extremely high precision rates with practically no false alarms. Fig. 6.21 shows the face detection and alignment process.

To estimate the 3D scene location of an aligned face from its image coordinates and size, it is necessary to update the intrinsic and extrinsic parameters of moving cameras at every point in time. The parameters are updated using the actual pan, tilt and zoom values read from the cameras through their RS-232 serial interface. An initial calibration is performed for each camera in its rest position ($pan = tilt = 0^\circ$) using standard calibration techniques [109], yielding initial values for the camera location in the scene T_{init} and its base rotation

Figure 6.21: Face detection and alignment in active camera images. In a first pass, a frontal face is detected by scanning the region surrounding the person track. In a second pass, the inside of the face rectangle is scanned with an eye detector. If two eyes can be found, the face is aligned and passed on for recognition. The procedure guarantees extremely low false alarm rates

R_{init}, as well as focal length estimates at 9 discrete zoom steps $f_{x,0} \dots f_{x,8}$. The camera rotation matrix is then continuously updated from the latest pan and tilt information, by multiplying the initial rotation matrix with a "correction matrix" R_{corr} (Eq. 6.1),

$$R_{act} = R_{init} \cdot \begin{pmatrix} \cos(\beta) & \sin(\alpha)\sin(\beta) & -\cos(\alpha)\sin(\beta) \\ 0 & \cos(\alpha) & \sin(\alpha) \\ \sin(\beta) & -\sin(\alpha)\cos(\beta) & \cos(\alpha)\cos(\beta) \end{pmatrix} \tag{6.1}$$

with α the read camera pan angle and β the tilt angle.

The focal length itself is not directly readable and is interpolated for the current camera zoom step from the discrete values $f_{x,0}$ to $f_{x,8}$, using a 4th order polynomial function. Even though interpolation introduces some imprecision, the maximum observed deviation error comprised only a few pixels, which is completely sufficient for our purpose.

Using the up to date intrinsic and extrinsic camera parameters, and assuming a standard eye distance of 7cm, the 3D location of each aligned and identified frontal face is computed.

Face Recognition

The face recognition algorithm is the same as described in Section 3.3.2. For training, feature vectors were obtained by automatically capturing sample images for known subject at different points in the room using the active cameras, and applying the same detection and alignment techniques described above. Training was done offline using roughly 70-180 images per person.

6.6.2 Speech Detection and Recognition

In parallel to face identification, speech localization and speaker identification is performed on a separate machine. For localization, the four T-shaped microphone arrays installed on the room walls are used. For speaker identification, only one channel from a table-top microphone is used. Three subtasks are accomplished as follows:

- Speech detection and segmentation: In the live system, this is done by thresholding the power spectrum of the table top audio signal. Speech segments of more than 1 second length are extracted and fed to the identification module.
- Speaker Identification: The algorithm for speaker ID is the same as described in Section 3.3.3. Speakers are modeled using a 32-component Gaussian Mixture Model (GMM). The inputs to the GMMs are the MFCC coefficients computed on the segmented speech from the table top channel. For each speaker, one GMM is trained offline on a 30 second speech segment recorded using the same table top microphone. The recognition itself is made on segments of 1 to 5 seconds, with longer segments being broken down into smaller ones, to allow for intermediate identification results and avoid the faulty inclusion of speech from multiple speakers into a lengthy segment. Cepstral mean subtraction and feature warping are performed on the audio signal to reduce channel noise and reverberation effects.
- Speech Source Localization: In contrast to the offline system, localization in the live system is performed by simple Kalman filtering of the GCC-computed time delays between the various microphone pairs. The details of the Kalman filter source localizer can be found in [41]. It has been evaluated in the CLEAR 2006 evaluations and reached somewhat lower scores than the JPDAF approach in the single person tracking case, with 66%$MOTA$, as compared to 78% in the JPDAF case. The reason the Kalman filter is used in the live system is realtime performance, as accuracies are largely sufficient for demonstration purposes.

6.6.3 Target and Camera Selection Strategies, Recognition of Standard Events

Based on the actual configuration of persons in the scene and on the recognition confidence for each, persons of interest are determined and the best active cameras configuration for their observation selected. A variety of camera and target selection strategies are conceivable, of which a few have been experimented with:

- Switching attention of all cameras to the currently active speaker: This strategy assumes speakers are the most important actors and puts the priority on achieving high recognition rates for them first.
- Achieving high identification confidence for all room occupants as fast as possible. This would prioritize participants that have not been identified yet.
- Trying to refresh all identities of all participants as regularly as possible. This is a good strategy if e.g. the confidence in the tracker's accuracy is low.
- Focusing two cameras on one person increases the chances of getting a frontal face. Alternatively, split cameras among users, possibly choosing the best camera for a user based on head orientation estimates.
- In situations where a main speaker can be clearly identified, keep one camera on the speaker and use the others to examine the audience.
- Define regions of high priority in the room, e.g. the door, to quickly identify new persons entering the room, etc.

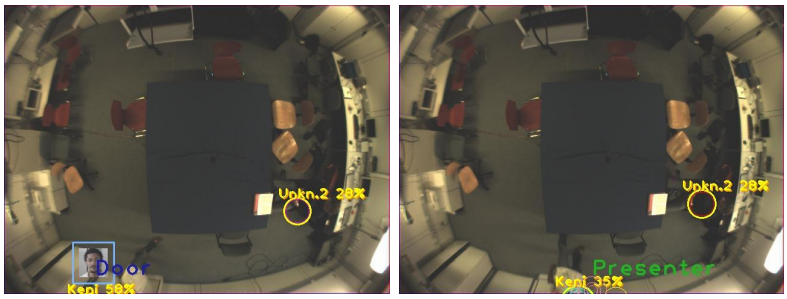
The currently implemented target and camera selection strategy does the following:

- It consecutively scans the locations of all participants using all active cameras simultaneously to increase the chances of capturing a frontal face. Targets of attention are switched regularly at 10 second intervals.
- Whenever a person track is near the entrance door, the active camera offering best views of the door area is immediately steered to capture the faces of eventual newcomers.
- Whenever a person is found to stay near the whiteboard for a certain period of time, another active camera, offering best views of the eventual presenter, is dedicated to following that person.

The recognition of simple events, such as a person entering the room, or a presenter staying near the whiteboard can be directly inferred from tracking and speech activity information. Figure 6.22 depicts the two recognized events "person at door" and "presenter" superimposed on the room's top view.

6.6.4 Experiments using the Live System

The integrated system was evaluated many times on various sample interaction scenarios involving several users entering the room and engaging in conversation in a meeting-like setting. The users were free to sit around the meeting table, occasionally stand up to give explanations in front of the whiteboard, walk



(a) **(b)**

Figure 6.22: Examples of standard events detected using the knowledge about person locations. In 6.22(a), the event of a person entering the room is recognized. In 6.22(b), the continuous movement of a person in front of the whiteboard area prompted the system to detect the start of a presentation.

around, etc. Figure 6.23 shows an example scenario evaluated using the live system. The database for identification comprised 7 to 10 users, with some users known purely through their faces and others only through their voice. The samples for training of the face recognizer were captured automatically over the course of several months, whereas training of voices was done on data from controlled recording sessions.

The ID tracking system was started as diverse points of simulated meetings, conversations, etc, to prevent it from capturing easy face snapshots, e.g. as users enter the room. The system automatically initiates and maintains tracks for 3 to 5 users, and gradually recognizes their identities in time. As opposed to the case of the offline CLEAR recordings, the visual identification modality is much more accurate, due to the captured high resolution face images. The main drawback of face recognition, even in the live system, is still the variation in head pose of the users, such that frontal faces for in some situations can rarely be observed. A greater camera coverage or a detection and identification on half profile views would substantially improve performance. Also, it was found that the detection of eyes in the close-up views is sometimes still problematic, such that other alignment techniques, such as e.g. Active Appearance Models (AAMs) [29] should be further investigated. As in the case of the offline evaluations, it was found that the acoustic localization and identification can be a very powerful tool in situations with low environmental noise and little crosstalk. The live implementation runs in realtime on standard hardware at the video framerate of 15 frames per second.

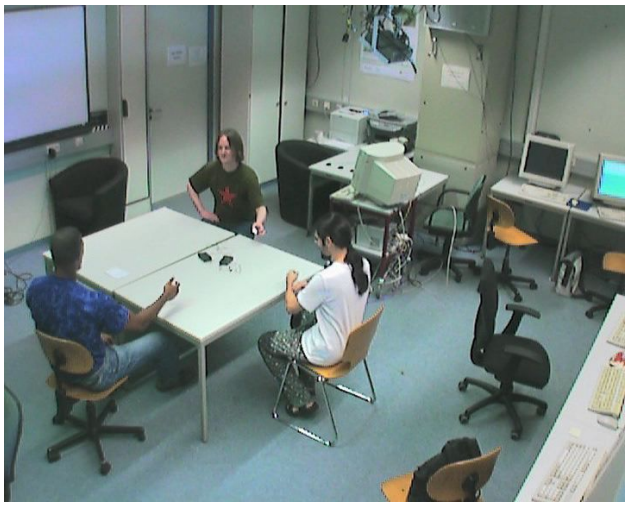


Figure 6.23: An example evaluation scenario for the live identity tracking system.

7 Conclusion

In this thesis, a novel framework was presented for the tracking and identification of multiple persons in a smart environment using distantly placed audio-visual sensors. The challenges addressed are the limited availability of reliable cues for person identification as well as the hard problems of tracking and data association in a noisy, cluttered and uncontrolled environment. The proposed approach, the *Joint Identity Tracking (JIT)* filter, builds on the opportunistic capture and joint probabilistic integration of tracking as well as face and voice identification features, gained from several cameras and microphone arrays, whenever these cues can be captured with a sufficient degree of confidence. It automatically detects persons at any location in the smart space and initializes tracks without the need for cooperation or explicit interaction. It builds person models on-the-fly, learns in and continuously adapts person-specific discriminative features in an unsupervised way. It adapts automatically to a wide variety of smart environments without the need for manual tuning. It estimates person locations audio-visually based on a variety of features using probabilistic filtering and robust Monte-Carlo approximations. It degrades gracefully in the presence of feature extraction errors, sensor or modality failure, and probabilistically associates face and voice observations to persons, even in the event of severe occlusion, low sensor coverage or missing person tracks. It derives identities jointly for multiple users based on the sequential probabilistic filtering of all available observations. It performs temporal and cross-modal fusion of identification results, based on the combination of normalized, warped confidence scores. It is designed to operate in the open set identification case, localizing and identifying subsets of known persons interacting with further unknown person and continuously adapts the confidences for all made identifications.

7.1 Summary and Discussion

A systematic evaluation procedure with associated performance metrics has been introduced to measure multiple person tracking and identity tracking accuracies. Using these metrics, the proposed *JIT* approach was systematically evaluated on a large database of audio-visual recordings featuring small meetings held in various smart rooms, the CLEAR Interactive Seminar database. Experimental results have shown:

- the advantage of fusing multiple modalities, for the unconstrained smart environment scenario, when using distantly placed sensors. Multimodal analysis not only allow to increase tracking and identification accuracies, it also exploits the true complementary nature of the acoustic and visual domains, especially in the case of noisy, infrequent observations, or sensor or modality failure. It was also shown that the fusion of asynchronous audio-visual identification features observed with highly varying frequency, can be achieved with success through appropriate modeling and adaptation schemes. A person tracking accuracy of 77% and an identity tracking accuracy of 81% could be achieved for the difficult conditions posed by the CLEAR Interactive Seminars.
- the advantage of confidence-based fusion and filtering of highly confident information. This is especially the case in the presence of many unknown persons, when the underlying classification algorithms are trained for the closed set scenario.
- the advantage of performing temporal fusion, even if automatic association of observations to persons has to be performed. The temporal fusion is made possible by the continuous tracking of persons, or management of untracked person models, until new ID observations become available. In this sense, it was shown that tracking improves the quality of identification, but also that identification improves the quality of tracking, e.g. through the reinitialization of corresponding, previously lost person tracks upon correlating ID feature observations.
- the benefits of integrating all information, support maps, detections, localized tracking cues and localized identification cues at the global level. This is with respect to the overall accuracies that can be reached, but also with respect to the system behavior as individual components fail. The perfomance of the *JIT* filter was shown to degrade gracefully, keeping high localization accuracies, and almost unchanged identification accuracies, even in the presence of severe failures of individual feature extraction steps.

A realtime implementation of the Joint Identity Tracking approach in a smart perceptual room was also presented, with processing distributed over a network of computers. The live system makes use of active steerable cameras to fully exploit the potential of face identification, keeps track of multiple identities evolving in the room and recognizes a few standard events, such as ongoing presentations, or newcomers entering the room.

For the presented quantitative evaluations, a set of novel metrics and a new definition of the identity tracking task, applicable to the general open set case, were presented and tested on the CLEAR Interactive Seminar scenario. The

proposed *MOT* metrics have already been applied with success in the international CLEAR evaluations, and are being employed in a growing number of publications from other research groups [47; 65; 56]. While the *MIT* metrics have not yet been employed outside of the frame of this thesis, they also show the potential for wide-spread application in the future: As steady technological advances are made in the field of multimodal, multi-user Human Computer Interaction, systems that perform open set identity tracking will no doubt become increasingly available. Similarly, large distributed monitoring systems, that keep track of many users in wide-spread environments (e.g. office buildings) become more and more feasible. As in the field of visual multiple target tracking today, benchmarking and formal evaluation of such identity tracking systems should play an increasingly important role in the future. In this case, the here proposed *MIT* metrics could represent a first effort toward standardization and systematic evaluation in those domains.

7.2 Future Directions

The identity tracking approach presented here still offers many possibilities for improvement or extension. Future work could include:

- Increasing the performance of face identification. Although the use of active cameras allows to circumvent the problem of low face resolutions, they come at the cost of an additional effort in sensor installation and control. Moreover, in many setups (such as offline surveillance recordings, for example), steerable cameras may just not be available. Therefore, further efforts to increase the reliability of detection and alignment of faces in distant views should be worthwhile. Also, for the recognition of low resolution faces, perhaps super-resolution methods could be investigated, in addition to temporal fusion, to increase the quality of obtained face models.

  For both the cases of fixed and steerable cameras, a major improvement would also be the accurate recognition based on non-frontal views. In unconstrained scenarios, the pose of observed faces constitutes a great challenge. The extension of a steerable camera system, e.g., to estimate head orientations and use pose-specific classifiers could bring a major improvement.

- The quality of speaker identification can also be improved, for example by using beamforming techniques, of by combining the outputs of several acoustic classifiers from different microphones. Another point would be to employ more powerful speaker diarization techniques, allowing the automatic generation of longer speech segments for identification. The

mapping of speaker ID results for long speech segments could then be made by comparing the short-time history of speech source locations with the history of person trajectories.

- Another direction could be the development of more sophisticated active camera steering techniques. A network of cooperating active cameras would no longer necessitate the aid of fixed cameras, reducing the amount of sensors needed. Especially considering the JIT filter's ability at managing unobserved tracks outside of the field of view of sensors, it is thinkable of using only one or two centrally placed steerable cameras that would periodically refresh their knowledge about the state of the world (the locations and identities of persons), for example based on criteria of saliency of observations or information gain provided by certain camera actions.

  In application areas such as videoconferencing, this may be very useful, as a greater number of participating persons may be managed while using fewer sensors for observation. A few active cameras would then cooperate in keeping the focus on the relevant persons (for example dominant or active speakers) while providing to the remote party realtime information about their identities, e.g. in form of speaker names, roles, affiliations, etc.

- A natural extension of the JIT approach would be its application to a network of rooms, equipped with varying types of sensors. Users could be tracked, e.g., in an office building with some rooms equipped only with single microphones for speaker identification, some others, such as meeting rooms, equipped with multiple cameras and microphone arrays, and some floors equipped only with low coverage camera networks. The flexibility of the approach with respect to the type of sensors used could be exploited, as it is not necessary to guarantee full coverage or a full sensor setup in each part of the space. Owing to the joint estimation procedure, the presence or absence of persons in specific parts of the space could be inferred from observations coming from other parts, using wholly different sensors.

- Similarly, the approach may be well suited for application in the domain of ubiquitous computing. Research in this domain to a large part involves determining the locations of several users in a larger area, using wearable devices, infrared or ultrasound sensors, RFID tags, or a variety of distributed sensors such as laser range finders, radars, cameras, and so forth [39]. Although the work in this thesis concentrates on sensors which do not require the explicit cooperation of users, the developed method is easily extensible to include the above mentioned sensors as well.

- One more extension would be to relax the requirement of pre-calibrated sensors. A smart environment could profit from the information given by visual identification cues gained in separate camera images, or azimuths

for identified speech segments in individual microphone arrays to infer a reliable association of observations across sensors and gradually learn the calibration parameters of the distributed sensor setup. Similarly, tracking information could be used to accumulate observations for yet unknown persons and automatically train new classifiers in an unsupervised manner. In this way, the missing information about the person identity in one modality could also be completed by knowledge from the other modality.

- Finally, the modeling of more knowledge about the smart space itself, e.g. the positions of chairs, tables, entrances, noise sources, windows, and so forth, could greatly improve the quality of detection and localization, leading also to a better inference of identities. This factor was still relatively neglected in the current work. Modeling of the environment could be made by recognizing and tracking objects and object states, in addition to humans. It could just as well come in the form of schedules, modeled user habits, or preferred user clothing, which could help adjust the priors for tracking and identification.

Acknowledgments

At this place I would like to extend some words of gratifude to some people without whom the work on my PhD thesis would not have been possible, or at least not as pleasant.

First of all I would like to thank my supervisor, Alex Waibel, for giving me the opportunity to work on this challenging and fascinating topic. It was a wonderful chance to start working in the frame of a big European project, in a compentent team, with such innovative and ambitious goals, and it definitely brought me a lot, both personally and for my career.

Some very special words of thanks go to Rainer Stiefelhagen, who has really inspired me for the PhD position in Karlsruhe back in 2003 and who has accompanied me throughout the course of the thesis with his invaluable advice and support. Thank you for again and again inciting me to push forward, for example to submit publications or to take the lead in some of the CLEAR evaluation tasks. Your belief in our capabilities, the incredible patience and calm shown towards me and all the other members of the team even in stressful situations, your ability to inspire team spirit has always motivated me and others to give our best.

Another very special thanks goes to Kai Nickel, whom I met at the start of my thesis and whom I have been sharing an office with since. Kai was always an unlimited source of knowledge in seemingly all domains and, in retrospect, probably the most reliable person I have ever known. Thanks for the countless stimulating or productive conversations, the many tips and bits of advice, the inspiring, amusing, musically inspired or often hilarious discussions, for being a good colleague and a good friend. The past 6 years would definitely not have been the same without this.

I would also like to extend my thanks to the many colleagues, students and friends who have worked with me in the CHIL project and beyond: Joachim who made every common CHIL meeting or trip a thing to remember, Matthias who was always good for a witful joke, Florian, Maria, John, Hazim, Michael, whom I have worked with for so long, Margit for always greeting me with a happy smile, Mika, Tobias, Annette, and so many more. It was a fantastic time and I will always keep great memories of it.

Finally, I want to thank my parents, though they were not directly involved in the thesis, for helping me acquire the talents that would prove invaluable for my work and also for my life: the constant desire for self-improvement, the ability to appreciate novelty and variety under all their forms, and the strength to persevere in the pursuit of one's goals.

Own Publications

[1] C. T. Aslan, K. Bernardin, and R. Stiefelhagen. Automatic Calibration of Camera Networks based on Local Motion Features. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, Marseille, France, October 2008. Best Paper Award.

[2] K. Bernardin, A. Elbs, and R. Stiefelhagen. Detection-Assisted Initialization, Adaptation and Fusion of Body Region Trackers for Robust Multiperson Tracking. In *IEEE International Conference on Pattern Recognition*, Hongkong, August 2006.

[3] K. Bernardin, A. Elbs, and R. Stiefelhagen. Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment. In *The Sixth IEEE International Workshop on Visual Surveillance (in conjunction with ECCV)*, Graz, Austria, May 2006.

[4] K. Bernardin, T. Gehrig, and R. Stiefelhagen. Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 70–81, Baltimore, MD, USA, May 8-11 2007. Springer.

[5] K. Bernardin and R. Stiefelhagen. Audio-Visual Multi-Person Tracking and Identification for Smart Environments. In *ACM Multimedia 2007*, Augsburg, Germany, September 2007.

[6] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing, Special Issue on Video Tracking in Complex Scenes for Surveillance Applications*, May 2008.

[7] K. Bernardin, R. Stiefelhagen, and A. Waibel. Probabilistic Integration of Sparse Audio-Visual Cues for Identity Tracking. In *ACM Multimedia 2008*, Vancouver, B.C., Canada, October 2008.

[8] K. Bernardin, F. van de Camp, and R. Stiefelhagen. Automatic Person Detection and Tracking using Fuzzy Controlled Active Cameras. In *The Seventh IEEE International Workshop on Visual Surveillance (VS2007)*, Minneapolis, USA, June 2007.

[9] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Christoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet. The CHIL Audiovisual Corpus for Lecture and Meeting Analysis inside Smart Rooms. In *Language Resources and Evaluation*, number 41 in Springer, 2007.

[10] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan. The CLEAR 2006 Evaluation. In *Multimodal Technologies for Perception of Humans, Proceedings of the first International CLEAR evaluation workshop, CLEAR 2006*, number 4122 in Springer LNCS, pages 1–45, Southampton, UK, April 6-7 2006.

[11] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo. The CLEAR 2007 Evaluation. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 3–34, Baltimore, MD, USA, May 8-11 2007. Springer.

[12] R. Stiefelhagen, K. Bernardin, H. Ekenel, J. McDonough, K. Nickel, M. Voit, and M. Woelfel. Audio-visual perception of a lecturer in a smart seminar room. *Signal Processing - Special Issue on Multimodal Interfaces*, 86(12), 2006.

Bibliography

[1] AMI - Augmented Multiparty Interaction, http://www.amiproject.org.

[2] CAVIAR - Context Aware Vision using Image-based Active Recognition, http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[3] CHIL - Computers in the Human Interaction Loop, http://chil.server.de.

[4] CLEAR - Classification of Events, Activities and Relationships, http://www.clear-evaluation.org/.

[5] EEMCV - Empirical Evaluation Methods in Computer Vision, http://www.cs.colostate.edu/eemcv2005/.

[6] ETISEO - Video Understanding Evaluation, http://www.silogic.fr/etiseo/.

[7] The i-LIDS dataset, http://scienceandresearch.homeoffice.gov.uk/hosdb/cctvimaging-technology/video-based-detection-systems/i-lids/.

[8] PETS - Performance Evaluation of Tracking and Surveillance, http://www.cbsr.ia.ac.cn/conferences/VS-PETS-2005/.

[9] VACE - Video Analysis and Content Extraction, http://www.ic-arda.org.

[10] Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.

[11] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006.

[12] J. Ajmera, G. Lathoud, and L. Mccowan. Clustering and segmenting speakers and their locations in meetings. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04).*, volume 1, 2004.

[13] J. Ajmera and C. Wooters. A robust speaker clustering algorithm. *Proceedings of the IEEE Workshop on Automatic Speech Recognition Understanding*, 2003.

[14] D. Anguelov, K.-C. Lee, S. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–7, June 2007.

[15] J. Annesley, J. Orwell, and J. P. Renno. Evaluation of MPEG7 color descriptors for visual surveillance retrieval. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 105–112, 2005.

[16] Y. Bar-Shalom. *Tracking and data association*. Academic Press Professional, Inc., San Diego, CA,USA, 1987.

[17] C. Barras, X. Zhu, C.-C. Leung, J.-L. Gauvain, and L. Lamel. Acoustic speaker identification: The LIMSI CLEAR'07 system. In *Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, pages 233–239, 2008.

[18] H. Bay, T. Tuytelaars, V. Gool, and L. Surf: Speeded up robust features. In *9th European Conference on Computer Vision*, Graz Austria, May 2006.

[19] K. Bernardin, T. Gehrig, and R. Stiefelhagen. Multi-Level Particle Filter Fusion of Features and Cues for Audio-Visual Person Tracking. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 70–81, Baltimore, MD, USA, May 8-11 2007. Springer.

[20] K. Brady. MIT lincoln laboratory multimodal person identification system in the CLEAR 2007 evaluation. In *Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, pages 240–247. 2008.

[21] R. Brunelli, A. Brutti, P. Chippendale, O. Lanz, M. Omologo, P. Svaizer, and F. Tobia. A generative approach to audio-visual person tracking. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop*, pages 55–68, Southampton, UK, 2007. Springer LNCS 4122.

[22] C. Busso, S. Hernanz, C.-W. Chu, S.-I. Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S. Narayanan. Smart room: participant and speaker localization and identification. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, 2:1117–1120, 2005.

[23] C. Canton-Ferrer, J. Casas, and M. Pardas. Particle filtering and sparse sampling for multi-person 3d tracking. In *15th IEEE International Conference on Image Processing (ICIP 2008)*, pages 2644–2647, Oct. 2008.

[24] C. Canton-Ferrer, J. Salvador, J. Casas, and M.Pardas. Multi-person tracking strategies based on voxel analysis. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 91–103, Baltimore, MD, USA, May 8-11 2007. Springer.

[25] N. Checka, K. W. Wilson, M. R. Siracusa, and T. Darrell. Multiple person and speaker activity tracking with a particle filter. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, volume 5, pages V–881–4 vol.5, 2004.

[26] T. Chen, W. Yin, X. S. Zhou, D. Comaniciu, and T. Huang. Total variation models for variable lighting face recognition. volume 28, pages 1519–1524, Washington, DC, USA, 2006. IEEE Computer Society.

[27] Y. Chen and Y. Rui. Real-time speaker tracking using particle filter sensor fusion. In *Proceedings of the IEEE*, volume 92, pages 485–494, 2004.

[28] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal person recognition using unconstrained audio and video. In *Second Conference on Audio- and Video-based Biometric Person Authentication '99 (AVBPA '99)*, pages 176–181, Washington, DC, 1999.

[29] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.

[30] R. Cucchiara, C. Grana, G. Tardini, and R. Vezzani. Probabilistic people tracking for occlusion handling. In *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004*, volume 1, pages 132–135 Vol.1, 2004.

[31] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893, 2005.

[32] T. Darrell, D. Demirdjian, N. Checka, and P. Felzenszwalb. Plan-View trajectory estimation with dense stereo background models. In *Proceedings of the International Conference on Computer Vision*, pages 628–635, Vancouver, BC, July 2001.

[33] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 601, Washington, DC, USA, 1998. IEEE Computer Society.

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

[35] H. Ekenel and R. Stiefelhagen. Block selection in the local appearance-based face recognition scheme. In *CVPR Biometrics Workshop*, New York, USA, June 2006.

[36] H. K. Ekenel, Q. Jin, and M. Fischer. ISL Person Identification Systems in the CLEAR 2007 evaluations. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 256–265, Baltimore, MD, USA, May 8-11 2007. Springer.

[37] H. K. Ekenel and R. Stiefelhagen. Analysis of local appearance-based face recognition: Effects of feature selection and feature normalization. In *CVPR Biometrics Workshop*, New York, USA, June 2006.

[38] S. Fidler. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):337–350, 2006.

[39] V. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello. Bayesian filtering for location estimation. *Pervasive Computing, IEEE*, 2(3):24–33, July-Sept. 2003.

[40] D. Gatica-Perez, G. Lathoud, J.-M. Odobez, and I. McCowan. Audiovisual probabilistic tracking of multiple speakers in meetings. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(2):601–616, Feb. 2007.

[41] T. Gehrig and J. McDonough. Tracking multiple speakers with probabilistic data association filters. In *Multimodal Technologies for Perception of Humans, Proceedings of the First International CLEAR Evaluation Workshop*, Southampton, UK, 2007. Springer LNCS 4122.

[42] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660, 2001.

[43] P. W. Grosse, H. Holzapfel, and A. Waibel. Confidence based multimodal fusion for person identification. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 885–888, New York, NY, USA, 2008. ACM.

[44] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, and R. Bolle. Face cataloger: Multi-scale imaging for relating identity to location. In *AVSS '03: Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, page 13, Washington, DC, USA, 2003. IEEE Computer Society.

[45] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Who? when? where? what? a real time system for detecting and tracking people. In *FG '98: Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, page 222, Washington, DC, USA, 1998. IEEE Computer Society.

[46] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[47] K. Heath and L. Guibas. Multi-person tracking from sparse 3D trajectories in a camera sensor network. In *Second ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2008)*, pages 1–9, Sept. 2008.

[48] H.K.Ekenel and R. Stiefelhagen. A generic face representation approach for local appearance based face verification. In *CVPR IEEE Workshop on Face Recognition Grand Challenge Experiments*, San Diego, CA, USA, June 2005.

[49] H.K.Ekenel and R. Stiefelhagen. Local appearance based face recognition using discrete cosine transform. In *13th European Signal Processing Conference (EUSIPCO)*, Antalya Turkey, September 2005.

[50] K. S. Huang and M. M. Trivedi. Distributed video arrays for tracking, human identification, and activity analysis. In *Proceedings of the 2003 International Conference on Multimedia and Expo, 2003. ICME '03*, volume 2, 2003.

[51] S. S. Intille, J. W. Davis, and A. F. Bobick. Real-time closed-world tracking. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 697, Washington, DC, USA, 1997. IEEE Computer Society.

[52] M. Isard and A. Blake. Condensation–conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1), 1998.

[53] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, December 2005.

[54] Q. Jin, T. Schultz, and A. Waibel. Far-field speaker recognition. *Special Issue of IEEE Transactions on Audio, Speech & Language on Speaker and Language Recognition*, September 2007.

[55] T. Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE Trans. on Comm. Technology*, 15:52–60, Feb. 1967.

[56] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):319–336, Feb. 2009.

[57] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud, and R. Horaud. Detection and localization of 3d audio-visual objects using unsupervised clustering. In *IMCI '08: Proceedings of the 10th international conference on Multimodal interfaces*, pages 217–224, New York, NY, USA, 2008. ACM.

[58] Z. Khan, T. Balch, and F. Dellaert. Efficient particle filter-based tracking of multiple interacting targets using an MRF-based motion model. In *Proceedings. 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003).*, volume 1, pages 254–259 vol.1, 2003.

[59] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[60] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoustic Speech Signal Processing*, 24(4):320–327, August 1976.

[61] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer. Multi-camera multi-person tracking for EasyLiving. In *Proceedings of the third IEEE International Workshop on Visual Surveillance*, pages 3–10, 2000.

[62] O. Lanz. Approximate Bayesian Multibody Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1436–1449, September 2006.

[63] O. Lanz, P. Chippendale, and R. Brunelli. An appearance-based particle filter for visual tracking in smart rooms. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 57–69, Baltimore, MD, USA, May 8-11 2007. Springer.

[64] Y. Li, A. Dore, and J. Orwell. Evaluating the performance of systems for tracking football players and ball. In *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Como, Italy, September 2005.

[65] K.-C. Lien and C.-L. Huang. Multiview-based cooperative tracking of multiple human objects. volume 2008, pages 1–13, New York, NY, United States, 2008. Hindawi Publishing Corp.

[66] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *IEEE ICIP 2002*, volume 1, pages 900–903, September 2002.

[67] N. C. P. Lin, N. P. C. Tseng, and N. L. G. Chen. Nearly lossless content-dependent low-power dct design for mobile video applications. volume 0, pages 1238–1241, Los Alamitos, CA, USA, 2005. IEEE Computer Society.

[68] M. Liu, Y. Chen, X. Zhou, X. Zhuang, M. Hasegawa-Johnson, and T. Huang. Multichannel and multimodality person identification. pages 248–255, 2008.

[69] J. Luque, X. Anguera, A. Temko, and J. Hernando. Speaker diarization for conference room: The UPC RT07s evaluation system. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 543–554, Baltimore, MD, USA, May 8-11 2007. Springer.

[70] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. In *Seventh International Conference on Computer Vision (ICCV'99)*, volume 1, pages 572–578, 1999.

[71] J. MacCormick and A. Blake. Probabilistic exclusion and partitioned sampling for multiple object tracking. *Int. Journal of Computer Vision*, 39(1):57–71, 2000.

[72] J. MacCormick and M. Isard. Partitioned sampling, articulated objects, and interface-quality hand tracking. In *ECCV '00: Proceedings of the 6th European Conference on Computer Vision-Part II*, pages 3–19, London, UK, 2000. Springer-Verlag.

[73] A. Martinez and R. Benavente. The AR face database. Technical report, CVC Technical Report 24, June 1998.

[74] A. M. Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. In *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, volume 24, pages 748–763, 2002.

[75] S. J. McKenna and H. Nait-Charif. Tracking human motion using auxiliary particle filters and iterated likelihood weighting. *Image and Vision Computing*, 25(6):852–862, 2007.

[76] V. Menon, B. Jayaraman, and V. Govindaraju. Integrating recognition and reasoning in smart environments. In *Proc. IET 4th International Conference on Intelligent Environments*, pages 1–8, July 21–22, 2008.

[77] I. Mikic, K. Huang, and M. Trivedi. Activity monitoring and summarization for an intelligent meeting room. In *HUMO '00: Proceedings of the Workshop on Human Motion (HUMO'00)*, page 107, Washington, DC, USA, 2000. IEEE Computer Society.

[78] A. Milstein, J. N. Sanchez, and E. T. Williamson. Robust global localization using clustered particle filtering. In *AAAI'02*, pages 581–586, 2002.

[79] A. Mittal and L. S. Davis. M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. In *International Journal of Computer Vision*, volume 51, pages 189–203, February 2003.

[80] B. Moghaddam. Bayesian face recognition. *Pattern Recognition*, 33(11):1771–1782, November 2000.

[81] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Christoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet. The CHIL Audiovisual Corpus for Lecture and Meeting Analysis inside Smart Rooms. In *Language Resources and Evaluation*, number 41 in Springer, 2007.

[82] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, 5(1):32–38, March 1957.

[83] M. Nechyba, L. Brandy, and H. Schneiderman. Pittpatt face detection and tracking for the CLEAR 2007 evaluation. In *Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, number 4625 in Springer LNCS, Baltimore, MD, USA, 2007.

[84] A. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. Etiseo, performance evaluation for video surveillance systems. In *Proceedings of AVSS07*, pages 476–481, London, U.K., Sept 2007.

[85] K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough. A joint particle filter for audio-visual speaker tracking. In *ICMI '05: Proceedings of the 7th international conference on Multimodal interfaces*, pages 61–68, New York, NY, USA, 2005. ACM.

[86] K. Nickel and R. Stiefelhagen. Dynamic integration of generalized cues for person tracking. In *Proceedings of the European Conference on Computer Vision (ECCV'08)*, pages 514–526, 2008.

[87] M. Omologo and P. Svaizer. Acoustic event localization using a crosspower-spectrum phase based technique. In *Proc. ICASSP*, volume 2, pages 273–276, 1994.

[88] J. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Trans. Comput.*, 56(9):1189–1224, 2007.

[89] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*, pages 947–954, Washington, DC, USA, 2005. IEEE Computer Society.

[90] C. Podilchuk and X. Zhang. Face recognition using dct-based feature vectors. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, pages 2144–2147, Washington, DC, USA, 1996. IEEE Computer Society.

[91] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1):72–83, 1995.

[92] D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, volume 5, pages 953–956, 2005.

[93] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, 2003.

[94] A. Rudnicky, P. E. Rybski, S. Banerjee, and M. Veloso. Intelligently integrating information from speech and vision processing to perform light-weight meeting understanding. In *Proceedings of ICMI'05, the International Conference on Multi-modal Interfaces, International Workshop on Multimodal Multiparty Meeting Processing*, Trento, Italy, October 2005.

[95] A. Salah, R. Morros, J. Luque, C. Segura, J. Hernando, O. Ambekar, B. Schouten, and E. Pauwels. Multimodal identification and localization of users in a smart environment. *Journal on Multimodal User Interfaces*, 2(2):75–91, 2008.

[96] C. Segura, A. Abad, C. Nadeu, and J. Hernando. Multispeaker localization and tracking in intelligent environments. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 82–90, Baltimore, MD, USA, May 8-11 2007. Springer.

[97] K. Smith, S. Ba, J.-M. Odobez, and D. Gatica-Perez. Evaluating multi-object tracking. In *Workshop on Empirical Evaluation Methods in Computer Vision (EEMCV)*, San Diego, CA, June 2005.

[98] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. In *Int. Conference on Computer Vision - ICCV'07*, Rio de Janeiro, Brasil, October 2007.

[99] V. Stanford, J. Garofolo, O. Galibert, M. Michel, and C. Laprun. The NIST smart space and meeting room projects: signals, acquisition annotation, and metrics. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 4, pages IV–736–9 vol.4, 2003.

[100] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[101] A. Stergiou, A. Pnevmatikakis, and L. Polymenakos. The AIT multimodal person identification system for CLEAR 2007. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 221–232, Baltimore, MD, USA, May 8-11 2007. Springer.

[102] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, and P. Soundararajan. The CLEAR 2006 Evaluation. In *Multimodal Technologies for Perception of Humans, Proceedings of the first International CLEAR evaluation workshop, CLEAR 2006*, number 4122 in Springer LNCS, pages 1–45, Southampton, UK, April 6-7 2006.

[103] R. Stiefelhagen, K. Bernardin, R. Bowers, R. T. Rose, M. Michel, and J. Garofolo. The CLEAR 2007 Evaluation. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*, pages 3–34, Baltimore, MD, USA, May 8-11 2007. Springer.

[104] R. Stiefelhagen, K. Bernardin, H. Ekenel, J. McDonough, K. Nickel, M. Voit, and M. Woelfel. Audio-visual perception of a lecturer in a smart seminar room. *Signal Processing - Special Issue on Multimodal Interfaces*, 86(12), 2006.

[105] R. Stiefelhagen, R. Bowers, and J. Fiscus, editors. *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, volume 4625 of *LNCS*. Springer, Baltimore, MD, USA, May 8-11 2007.

[106] R. Stiefelhagen and J. Garofolo, editors. *Multimodal Technologies for Perception of Humans, First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR'06*. Number 4122 in LNCS. Springer, Southampton, UK, April 6-7 2006.

[107] S. Stillman and I. Essa. Towards reliable multimodal sensing in aware environments. In *PUI '01: Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–6, New York, NY, USA, 2001. ACM.

[108] H. Tao, H. S. Sawhney, and R. Kumar. A sampling algorithm for tracking multiple objects. In *ICCV '99: Proceedings of the International Workshop on Vision Algorithms*, pages 53–68, London, UK, 2000. Springer-Verlag.

[109] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3d machine visionmetrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, Aug. 1987.

[110] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part I*, pages 447–460, London, UK, 2002. Springer-Verlag.

[111] J. Vermaak, M. Gangnet, A. Blake, and P. Perez. Sequential monte carlo fusion of sound and vision for speaker tracking. In *Eighth IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 741–746 vol.1, 2001.

[112] J. Vermaak, S. J. Godsill, and P. Pérez. Monte carlo filtering for multitarget tracking and data association. *IEEE Transactions on Aerospace and Electronic Systems*, 41:309–332, 2004.

[113] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - CVPR 2001*, volume 1, pages 511–518, 2001.

[114] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), 1997.

[115] J. Wright, A. Ganesh, A. Yang, and Y. Ma. Robust face recognition via sparse representation. Technical report, University of Illinois, USA, 2007.

[116] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel. Multimodal people ID for a multimedia meeting browser. In *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 159–168, New York, NY, USA, 1999. ACM.

[117] F. Yin, D. Makris, and S. A. Velastin. Performance evaluation of object tracking algorithms. In *10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS2007), Rio de Janeiro, Brazil*, October 2007.

[118] W. You, S. Fels, and R. Lea. Studying vision-based multiple-user interaction with in-home large displays. In *HCC '08: Proceeding of the 3rd ACM international workshop on Human-centered computing*, pages 19–26, New York, NY, USA, 2008. ACM.

[119] W. Zajdel and B. J. A. Kröse. A sequential bayesian algorithm for surveillance with nonoverlapping cameras. *IJPRAI*, 19(8):977–996, 2005.

[120] Q. Zhao, J. Kang, H. Tao, and W. Hua. Part based human tracking in a multiple cues fusion framework. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 450–455, Washington, DC, USA, 2006. IEEE Computer Society.

[121] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.

[122] F. Ziliani, S. Velastin, F. Porikli, L. Marcenaro, T. Kelliher, A. Cavallaro, and P. Bruneaut. Performance evaluation of event detection solutions: the CREDS experience. In *Proc. IEEE Conference on Advanced Video and Signal Based Surveillance AVSS2005*, pages 201–206, September 2005.

[123] D. N. Zotkin, R. Duraiswami, and L. S. Davis. Joint audio-visual tracking using particle filters. *EURASIP J. Appl. Signal Process.*, 2002(1):1154–1164, 2002.