

Multimodal Audio-Visual Sensor Fusion for Multi-Object Tracking of Speakers in Indoor Environments

Referent: Jens-Peter Akelbein

Korreferent: Stephan Gimbel

Gliederung

1. MOTIVATION & FORSCHUNGSFRAGEN
2. VERWANDTE ARBEITEN
 - 2.1 RAUMAKUSTIK
 - 2.2 VISUELLE WAHRNEHMUNG
 - 2.3 SENSOR FUSION UND TRACKING
3. KONZEPTENTWICKLUNG
4. EXPERIMENTELLER AUFBAU
5. EVALUATION
6. ZUSAMMENFASSUNG
7. AUSBLICK
8. OFFENE DISKUSSION & FRAGEN

1. MOTIVATION & FORSCHUNGSFRAGEN

Motivation

Warum Sprecherverfolgung in Innenräumen?

- Zentrale Schlüsseltechnologie für „Smart Environments“
 - Gebäudesteuerung (z.B. Temperatur, Belüftung)
 - Intelligente Konferenzsysteme (z.B. Dynamische Audio und Video Regelung)
 - Menschliche Interaktion der Umgebung (z.B. Verkauf im Einzelhandel)

Limitationen einzelner Modalitäten

-  **Audio:** empfindlich gegenüber Nachhall, Lärm, Mehrdeutigkeit
 -  **Video:** abhängig von Sichtlinie, Beleuchtung, Verdeckung
- Motivation für multimodale Sensorfusion

Motivation

Zentrales Forschungsproblem

- Mangel an realistischen audio-visuellen Datensätzen
- Aufwendige Erhebung von Ground-Truth Daten
- Herausforderungen im Bereich Raumakustik, Sensorkalibrierung & Synchronisation

Zielsetzung dieser Arbeit

- Entwicklung einer reproduzierbaren audio-visuellen Simulation, räumlichen Sprecherverfolgung und Evaluation
- Systematischer Vergleich der Tracking-Leistung der Modalitäten (1) Audio (2) Video, (3) Audio-Video Fusion

Forschungsfragen

F1: Wie lassen sich realistische und reproduzierbare audio-visuelle Sensordaten für Szenarien mit mehreren Sprechern in Innenräumen simulieren?

F2: Wie kann eine Feature-Level-Sensorfusion entworfen und umgesetzt werden, um Audio- und Videoinformationen effektiv zu kombinieren?

F3: In welchem Ausmaß verbessert audio-visuelle Sensorfusion die Genauigkeit und Robustheit gegenüber unimodalem Tracking?

2. VERWANDTE ARBEITEN

2.1 RAUMAKUSTIK

Synthetische Audiodatengenerierung

Virtuelle Akustik-Simulationstools

Table 2.1: A comparison of maintained virtual acoustics simulation tools

Tool	Integration	Sources	Receivers	Simulation Method	Computation	Domain	Pricing
Pyroacoustics [16]	Python	Static (Multiple)	Static (Multiple)	ISM	Offline	Research	Free
GSound-SIR [17]	Python	Static (Multiple)	Moving (Single)	Ray Tracing	Offline	Research	Free
SoundSpaces 2.0 [18]	Python	Static (Multiple)	Moving (Single)	Ray Tracing	Offline	Research	Free
gpuRIR [5]	Python	Moving (Multiple)	Moving (Multiple)	ISM (with GPU acceleration)	Real-time	Research	Free
Virtual Acoustics [19]	Python, Unity, UE	Moving (Multiple)	Moving (Single)	FDTD	Offline	Research	Free
RAVEN [20]	Python, Unity, UE	Moving (Multiple)	Moving (Single)	Hybrid (ISM + Ray Tracing)	Offline	Research	Free (Research)
Meta XR Audio SDK [21]	Unity, UE	Moving (Multiple)	Moving (Single)	Ray Tracing	Real-time	Gaming	Free
Wwise + Reflect Plugin [22]	Unity, UE	Moving (Multiple)	Moving (Single)	Hybrid (ISM + Ray Tracing)	Real-time	Gaming	Commercial
Treble [23]	Standalone, Python	Static (Multiple)	Static (Multiple)	Hybrid (Ray Tracing / Radiosity + FEM)	Real-time	Room Optimization	Commercial

Weitere Vorteile:

- Benutzerfreundlichkeit
- Einfache Anpassung
- Etabliert in der Forschung

Positionsbasierte Schallquellenlokalisierung

3D Sound Source Localization (3D-SSL)

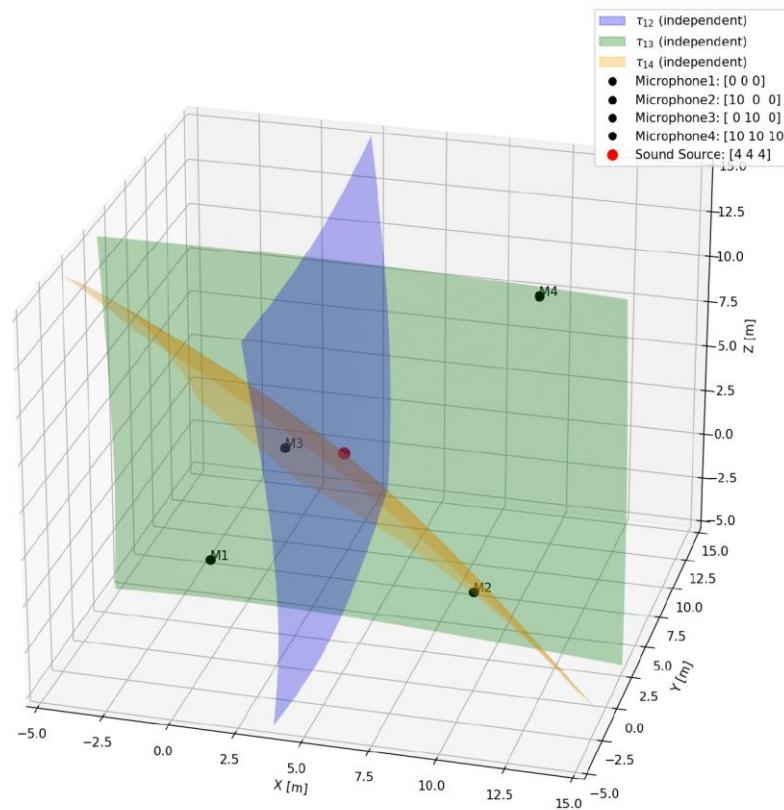


Figure 2.5: The 3D position of a single sound source is derived using multilateration of four microphones.

Eigenschaften:

- Lokalisation im Nahfeld
- Nicht-koplanare Mikrofonanordnung
- Time-Difference-of-Arrival (TDOA)
- Lokalisation einer einzelnen Schallquelle

2.2 VISUELLE WAHRNEHMUNG

Synthetische Bildatengenerierung

3D-Grafik-Engine-basierte Verfahren

Dataset	Simulator	Camera Projection	#Frames	3D	Pose	Segmentation	Depth
VIPER [51]	RAGE	Perspective	254k	✓	—	✓	—
GTA [52]	RAGE	Perspective	250k	—	—	✓	✓
JTA [49]	RAGE	Perspective	460k	✓	✓	—	—
MOTSynth [50]	RAGE	Perspective	1,382k	✓	✓	✓	✓
THEODORE [41]	Unity	Fisheye	100k	—	—	✓	—
THEODORE+ [46]	Unity	Fisheye	50k	✓	✓	✓	—

Table 2.3: A comparison of labeled large-scale synthetic video datasets.

Weitere Vorteile:

- Funktionserweiterung mit C# Skripten und HLSL Shadern
- Forschung zeigt, dass Machine-Learning-Modelle auf synthetischen Fisheye-Bildern generalisieren (Simulation-to-Reality Gap)

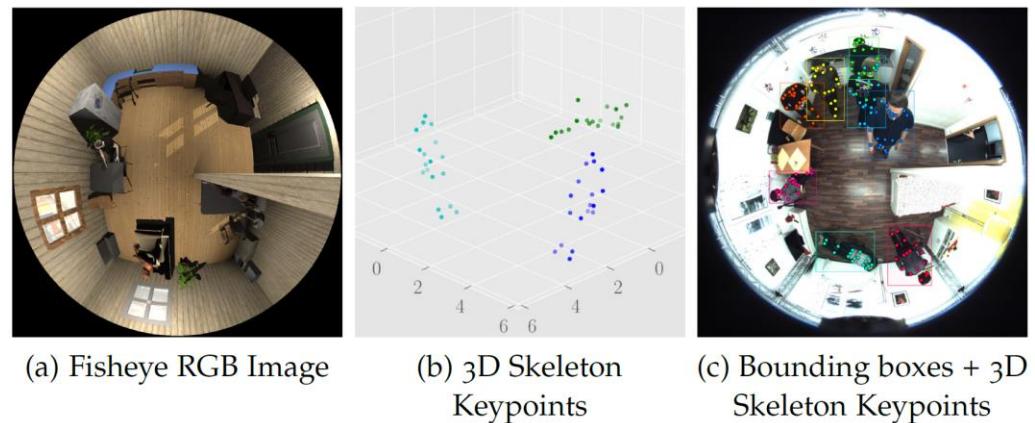


Figure 2.7: The THEODORE+ dataset as an example for a synthetic, annotated, top-down fisheye video dataset using the graphics engine *Unity* [46].

Kamerakalibrierung und Linsenverzeichnungskorrektur

Perspektivische Projektion

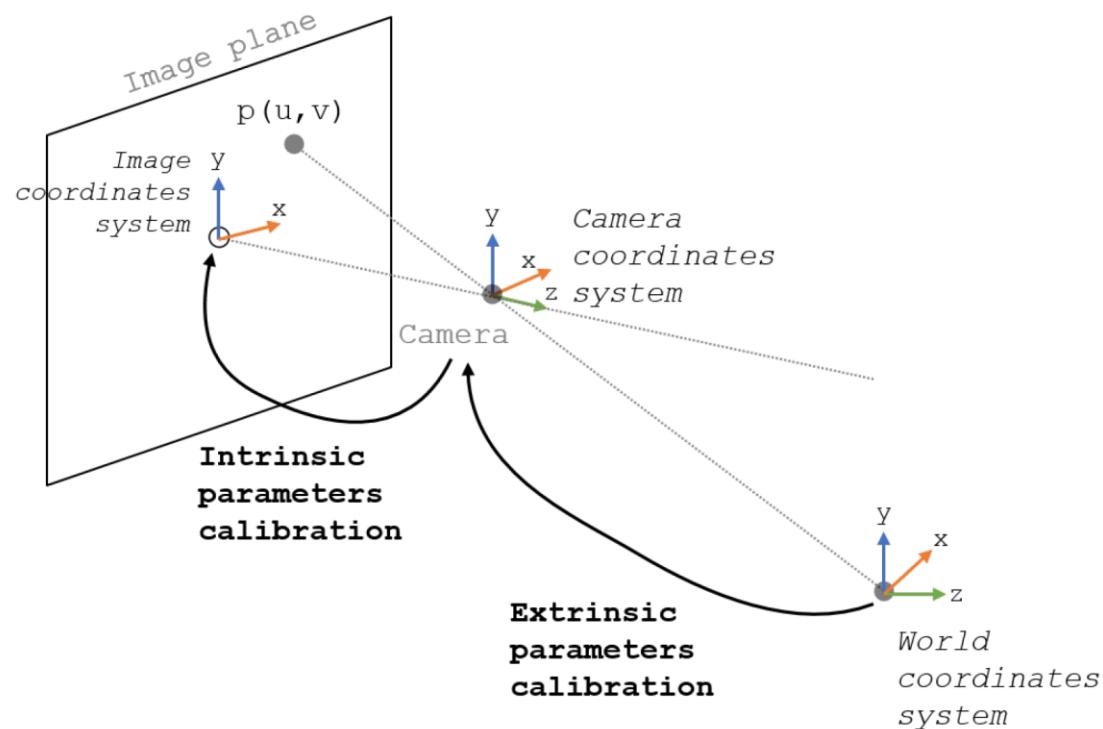


Figure 2.8: A schematic representation of the pinhole camera model. Its perspective projection is described by extrinsic and intrinsic parameters [54].

Geometrische Kamerakalibrierung:

- Extrinsischen Parameter: Position und Ausrichtung der Kamerarelativ zum Weltkoordinatensystem
- Intrinsischen Parameter: Optischen Eigenschaften der Kamera beschreiben die Projektion von Strahlen auf den Bildsensor.

Fisheye Projektionen:

- Kannala-Brandt Projektion (bis 180° diagonaler Blickwinkel)
 - Brown-Conrady Projektion (bis ~170° diagonaler Blickwinkel)
- Mathematisch modelliert als Polynom des Winkels zwischen dem Punkt und der Hauptachse.

Deep-Learning-basierte Objekterkennung

You Only Look Once (YOLO)

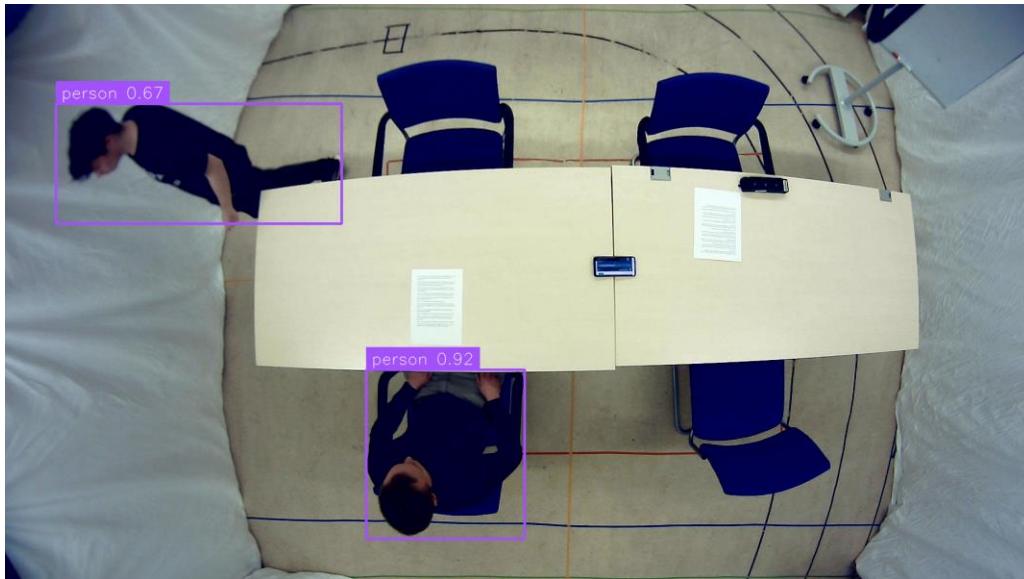


Figure A: Deep-learning object detection of two people in the laboratory.

Eigenschaften

- Once-Stage Detektor
- Detektion von mehreren Objekten als Bounding Boxes
- Hohe Genauigkeit, Robustheit & Echtzeitfähigkeit

Erkenntnisse aus der Vorarbeit:

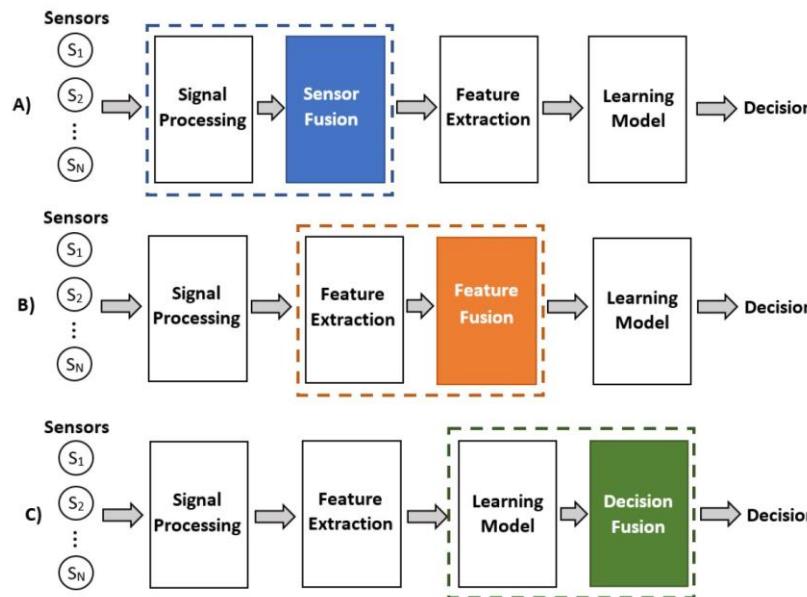
- Eignet sich zur Detektion von Personen
- Generalisiert gut auf Weitwinkel-Kamerabildern
- Beste Balance zwischen Genauigkeit und Geschwindigkeit

2.3 SENSOR FUSION UND TRACKING

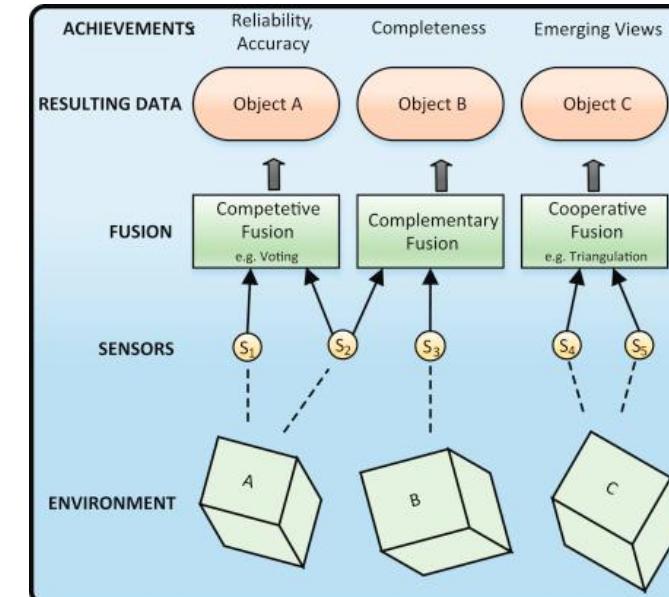
Definition und Kategorisierung

Definition: „Sensorfusion ist die Kombination von Sensordaten oder aus Sensordaten abgeleiteten Daten auf eine Weise, dass die resultierenden Informationen in gewisser Weise besser sind als die Informationen, die möglich wären, wenn diese Quellen einzeln verwendet würden.“ [68]

Kategorisierung nach Fusionsebene



Kategorisierung nach Sensorkonfiguration



[B]

[C]

Multi-Object Tracking

Übersicht von State-of-the-Art Algorithmen

Table 2.4: Overview and comparison of multi-object tracking algorithms.

Algorithm	Detection Paradigm	Processing Mode	Inference Strategy	Solution Nature
SORT [79]	TBD	Online	Local	Deterministic
DeepSORT [80]	TBD	Online	Local	Deterministic
ByteTrack [81]	TBD	Online	Local	Deterministic
FairMOT [82]	JDT	Online	Learned	Deterministic
MOTR [83]	JDT	Offline	Learned	Deterministic
TrackFormer [84]	JDT	Offline	Learned	Deterministic
MS-GLMB [85]	TBD	Online	Probabilistic Filtering	Stochastic
HybridTrack [78]	JDT	Online	Local + Learned	Stochastic

Weitere Vorteile:

- Unterstützt eine beliebe Anzahl von Sensoren bzw. Detektoren
- Tracking von Features im Euklidischen Raum
- Forschung zeigt erfolgreiche Anwendung auf 3D Tracking von Personen

Multi-Object Tracking

“Multimodale Raum-Zeit Permutations Problem”

= Unklare Zuordnung von Detektionen unterschiedlicher Modalitäten in Raum und Zeit

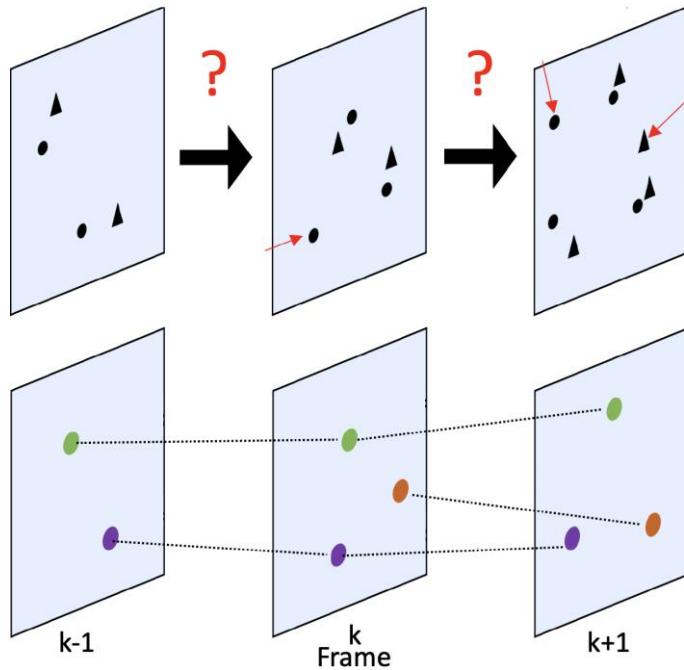


Figure D: Multi-modal space-time permutation problem.

Lösung:

1. **Sensor Fusion auf Feature-Ebene:** Überführung von Detektionen in einen einheitlichen Darstellungsraum
2. **Datenassoziation:** Identifizierung und korrekte Zuordnung von Detektionen aus beiden Modalitäten zu den jeweiligen Sprechern.
3. **Tracking:** Vorhersage von Sprecher Tracks im Zeitverlauf unter Berücksichtigung neuer oder verschwindender Sprechern, sowie verrauschte und fehlende Sensordetektionen

Multi-Object Tracking

Multi-Sensor Generalized Labeled Multi-Bernoulli Filter (MS-GLMB)

Grundlegende Funktionsweise:

1) Kalman-Filter mit Bewegungsmodell:

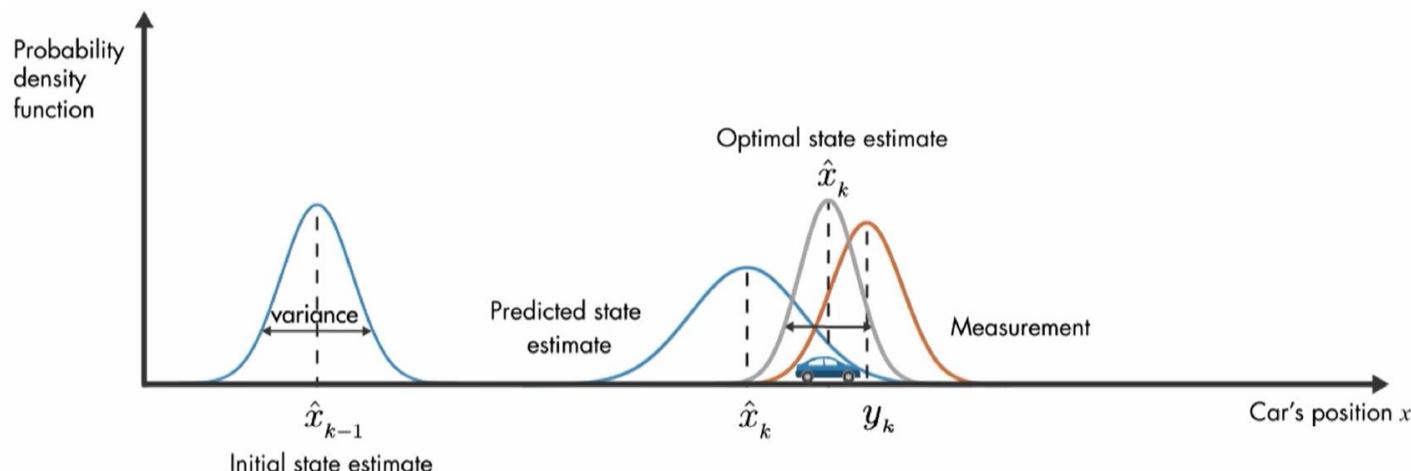


Figure E: Kalman filter example for a moving object.

- 2) Stochastische Approximation von Datenassoziationen
- 3) Adaptive Geburt basierend auf Gaußschen Wahrscheinlichkeiten

Multi-Object Tracking Evaluation

Higher Order Tracking Accuracy (HOTA)

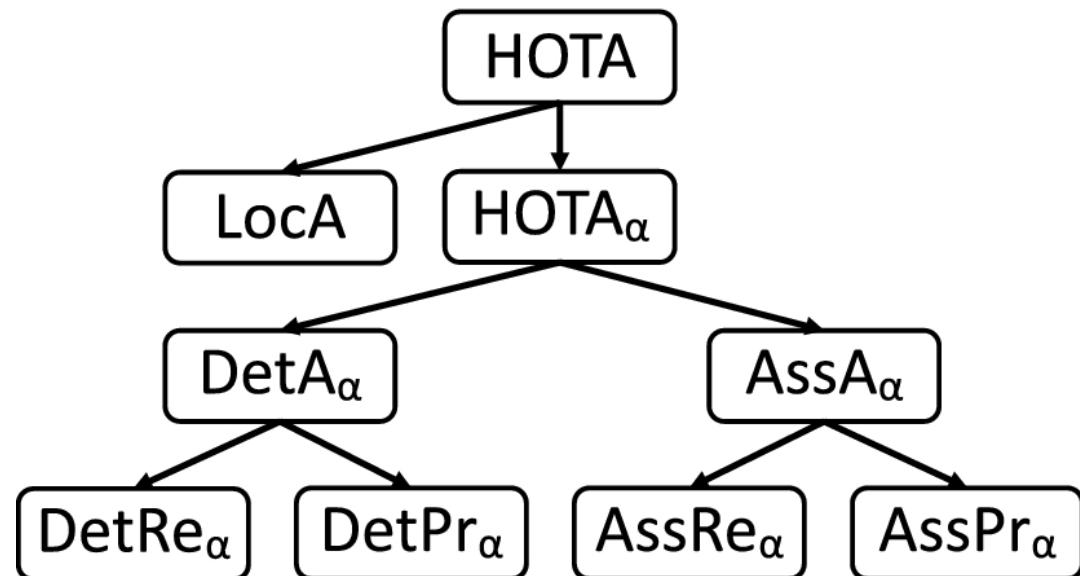


Figure F: HOTA metric and its sub-metrics with values between 0 and 100.

3. KONZEPTENTWICKLUNG

Entwurf einer Audio-Visuellen Softwarearchitektur

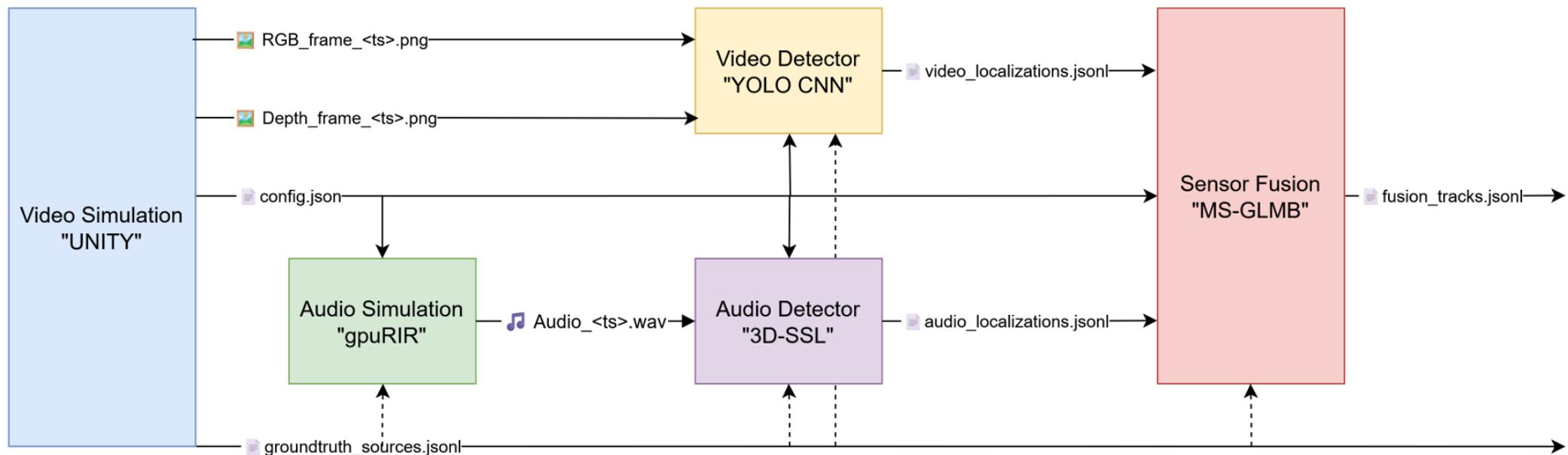


Figure 3.1: The proposed modular audio-visual software architecture design.

Forschungshypothesen

H1: Audio MOT

- Stark abhängig von Sprechaktivität & Raumhall
- Viele fehlende Detektionen und Track-Splits
- Erwartung: niedrigste HOTA-Scores

H2: Video MOT

- Hohe Genauigkeit bei Sichtkontakt
- Leistungsabfall bei Verdeckung
- Erwartung: hohe, aber instabile HOTA-Scores

H3: Audio-Video MOT

- Kombiniert räumliche Präzision & Robustheit
- Weniger Detektionslücken, stabilere Tracks
- Erwartung: höchste HOTA-Scores

4. EXPERIMENTELLER AUFBAU

Raumakustik Simulation

Simulation von Mikrofonen

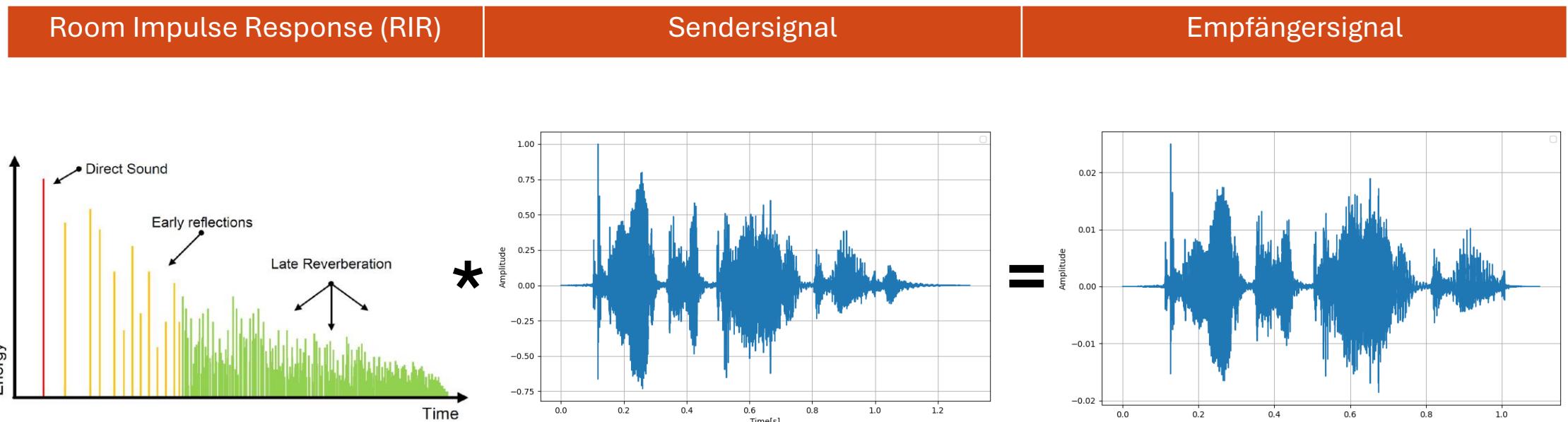


Figure 2.3: Simplified Room Impulse Response of a sound source and receiver in a shoebox shaped room [8].

Raumakustik Simulation

Simulation von Mikrofonen und sich räumlich bewegenden Sprechern

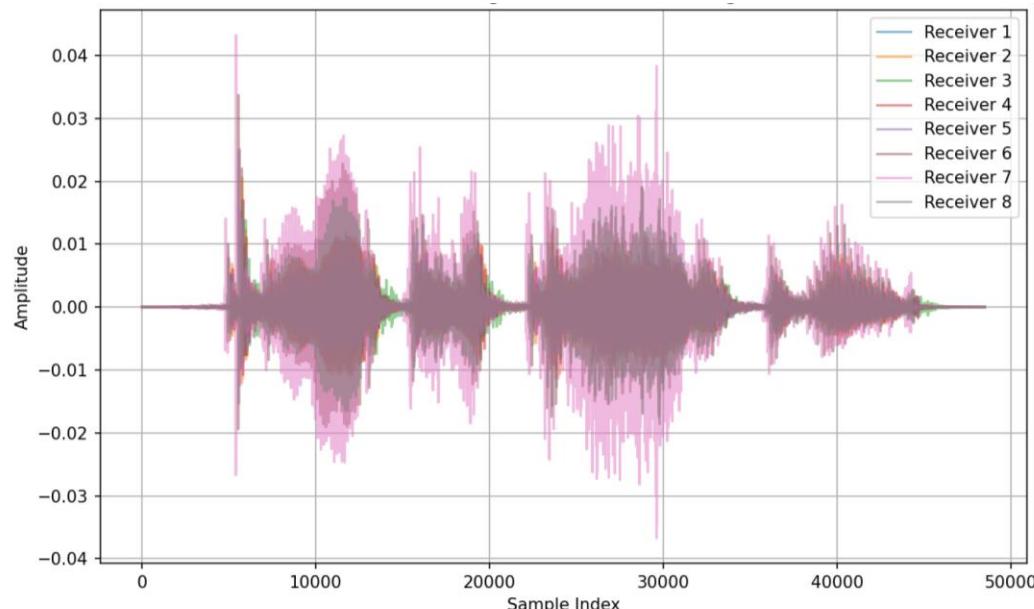


Figure 4.1: Synthetic microphone signals of a single moving male speaker generated with gpuRIR.

Raumakustische Modellierung

- Männliche Sprache („Can you keep a secret?“)
- Raumhallzeit $T60 = 0.35$ s (Büroräume nach DIN 18041)
- Segmentweise Faltung mit Simulationstool „gpuRIR“
→ Simulation bewegliche Sprecher
- Sprecher Laufgeschwindigkeit von 1 m/s
- Simulierter Frequenzbereich 200Hz – 20kHz
- Realistische Simulation von Geometrischer Akustik ab **146Hz** (Schröderfrequenz für $V=65\text{m}^3$ und $T60 = 0.35\text{s}$)

Erweiterung:

- Interpolation zwischen diskreten Sprecherpositionen (Positionsauflösung <1cm)

Audio Detektor

Sound Event Detection und Positionsbasierte Schallquellenlokalisierung

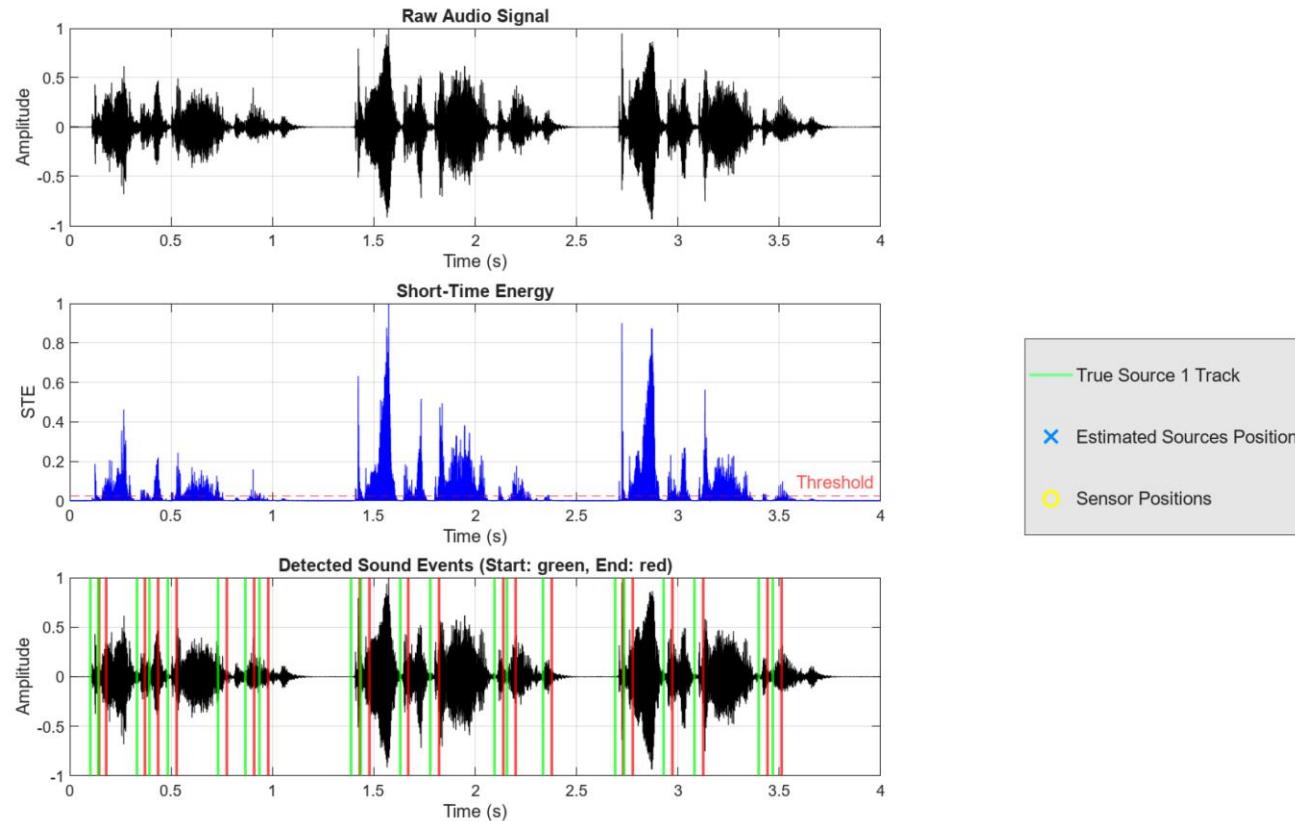


Figure 4.2: Processing steps of sound event detection: raw audio signal (top), short-time energy (center) and detected events (bottom).

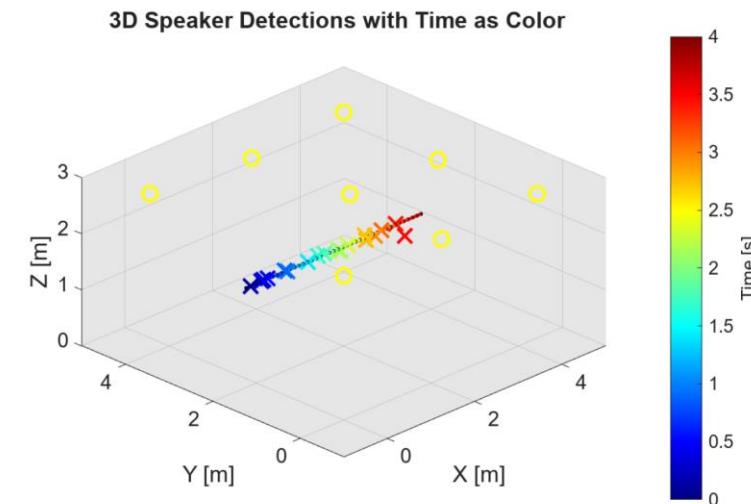


Figure 4.3: 3D-SSL detections of a moving speaker in a very low reverberation indoor environment ($T_{60} = 0.1$ s).

Visuelle Simulation

Simulation der Weitwinkelkamera

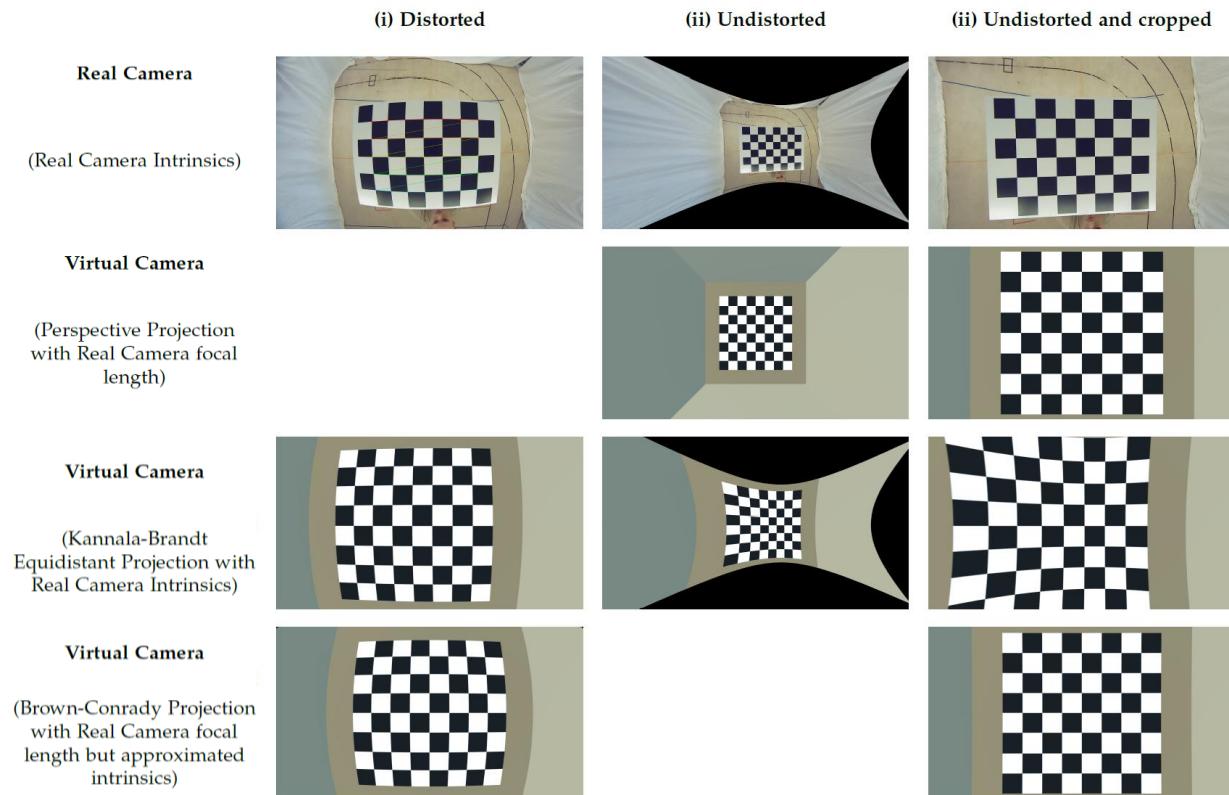


Figure 4.5: Visual comparison of real and virtual camera images after (i) distortion, (ii) undistortion and (iii) undistortion and cropping.

Kannala–Brandt Projektion:

- unzureichende Vorwärtsprojektion
- realistische Rückprojektion

Brown–Conradty Projektion

- approximierte Vorwärtsprojektion
- sehr präzise Rückprojektion

→ Entscheidung: Brown–Conradty

Visuelle Simulation

Definition und Simulation einer Raumszene



Figure 4.6: The indoor environment with three speakers simulated in Unity.



Figure 4.7: The synthetic fisheye camera image of the simulated scene

Video Detektor

Objekterkennung und 2D-3D Rückwärts-Projektion



Figure 4.8: Bounding box detections and point abstractions in the synthetic RGB fisheye camera image using the YOLO model.



Figure 4.9: Bounding box detections and point abstractions in the synthetic Depth fisheye camera image using the YOLO model.

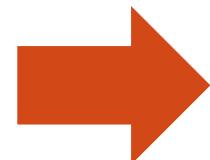
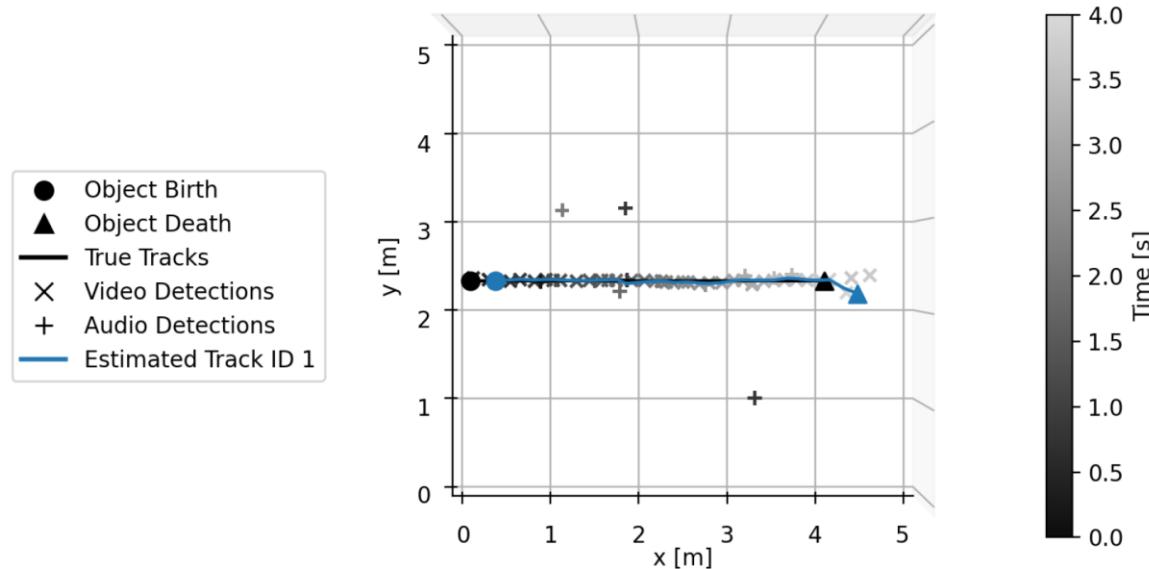


Figure 4.10: Undistorted 3D video detections of speakers in the indoor environment.

Audio-Visuelle Sensor Fusion und Tracking

Multi-object Tracking von sich räumlich bewegenden Sprechern



Unterschiedliche Positionen als Features:

- Audio Detektor: 3D Mund bzw Kopfzentrum
 - Video Detektor: 3D Mittelpunkt eines Sprecherdetektion
- Lösung: Orthogonale Projektion von 3D zu 2D

Figure 4.11: MS-GLMB speaker tracking results for tracking a single moving speaker using the audio and video detections.

5. EVALUATION

Simulierte Szenarien und Modalitäten

Szenarien:

S1: 1 Person läuft und spricht kontinuierlich

S2: 1 Person läuft und spricht kontinuierlich, ist jedoch visuell temporär verdeckt

S3: 3 Personen laufen. Die mittlere Person entgegengesetzt und spricht kontinuierlich

S4: 3 Personen laufen. Die mittlere Person entgegengesetzt. Sie sprechen sequentiell 1/3 der Zeit

Modalitäten:

M1: Audio → Datensatz MOT25A

M2: Video → Datensatz MOT25V

M3: Audio und Video → Datensatz MOT25AV

Evaluationsmetrik:

- Higher Order Tracking Accuracy (HOTA)

Szenario S1: 1 Person läuft und spricht kontinuierlich



Szenario S1: 1 Person läuft und spricht kontinuierlich

Modalität M1: Audio

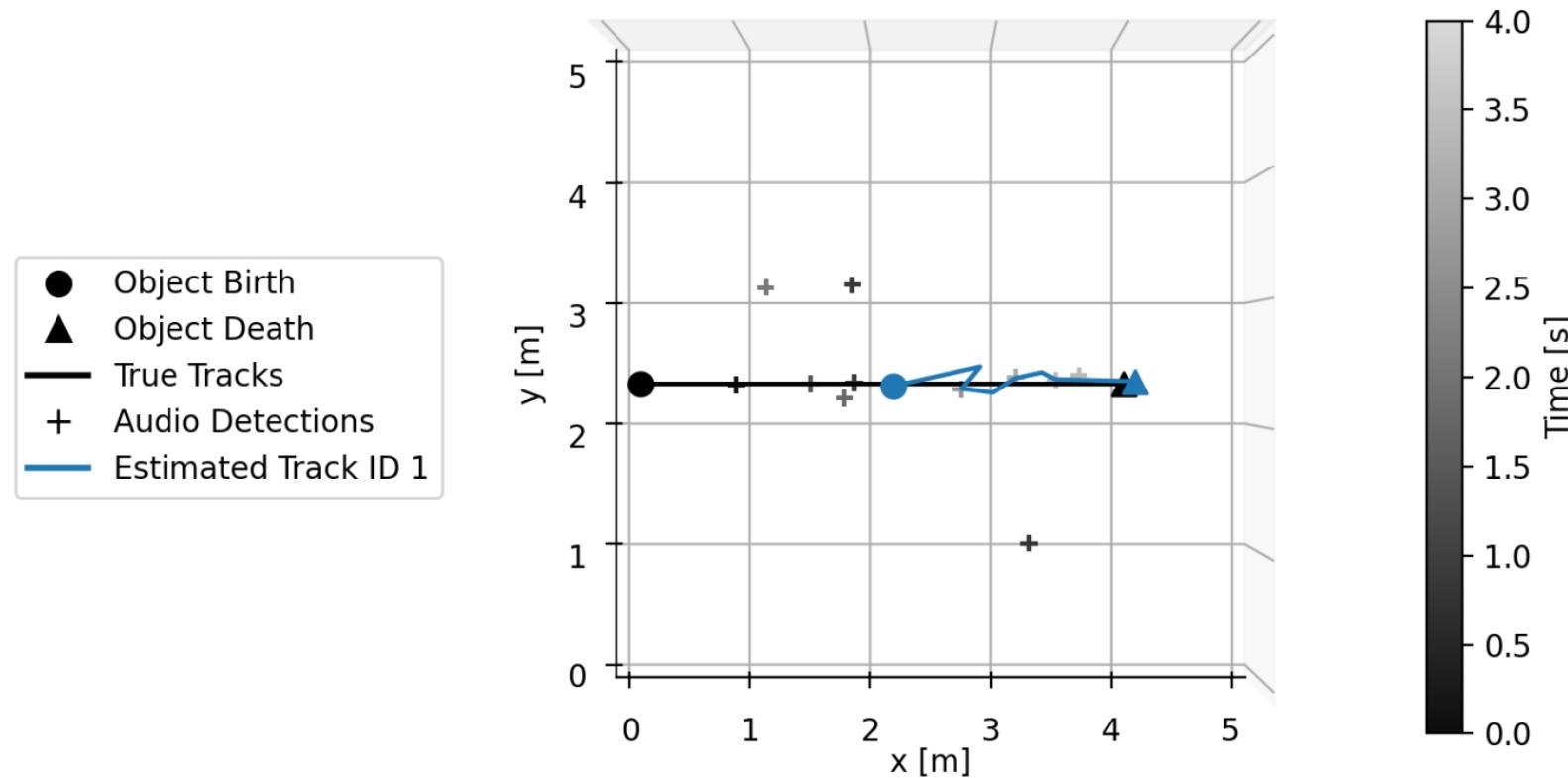


Figure 5.1: Tracking results of scenario S1 and modality M1.

Szenario S1: 1 Person läuft und spricht kontinuierlich

Modalität M2: Video

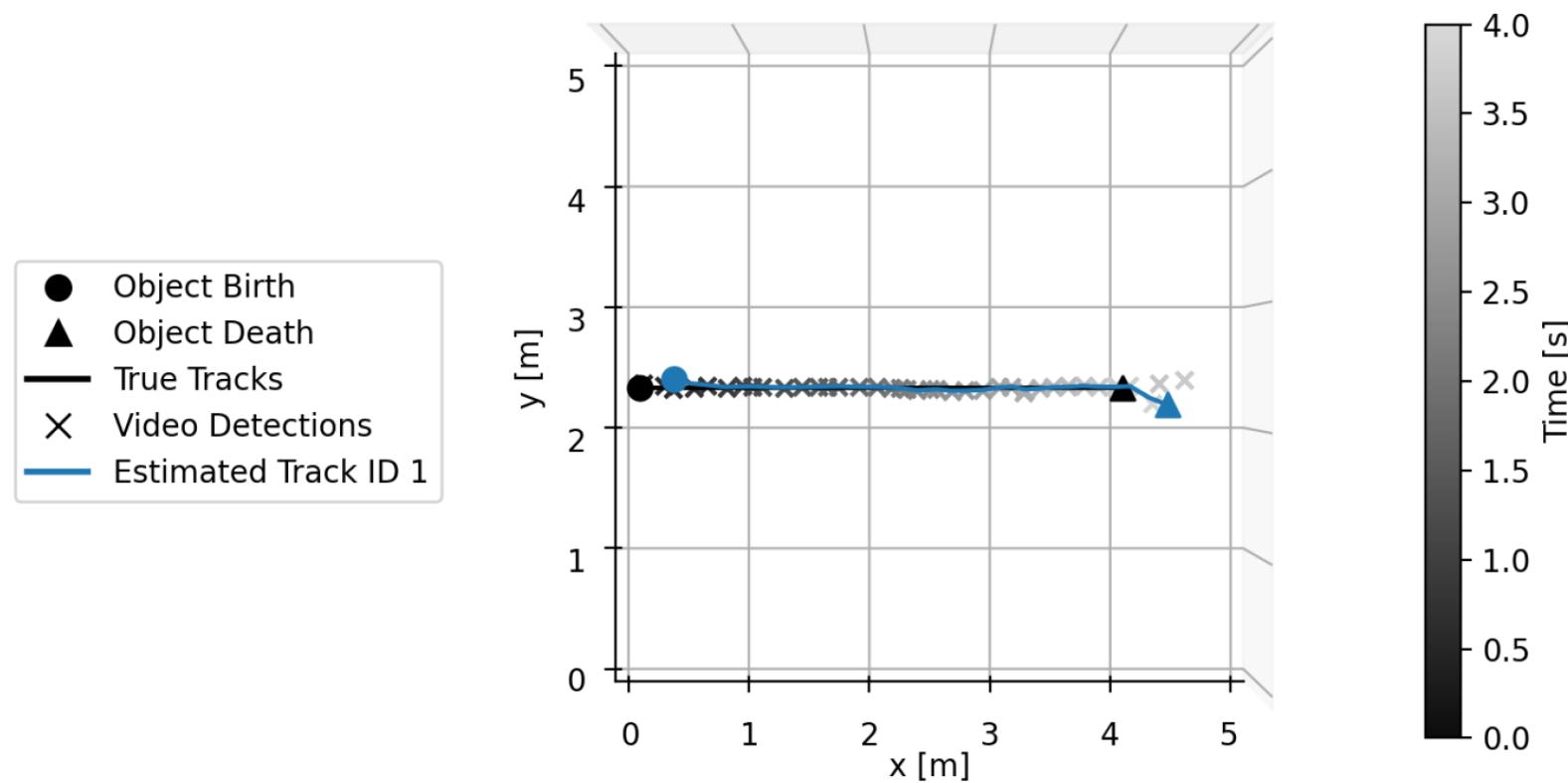


Figure 5.2: Tracking results of scenario S1 and modality M2.

Szenario S1: 1 Person läuft und spricht kontinuierlich

Modalität M3: Audio und Video

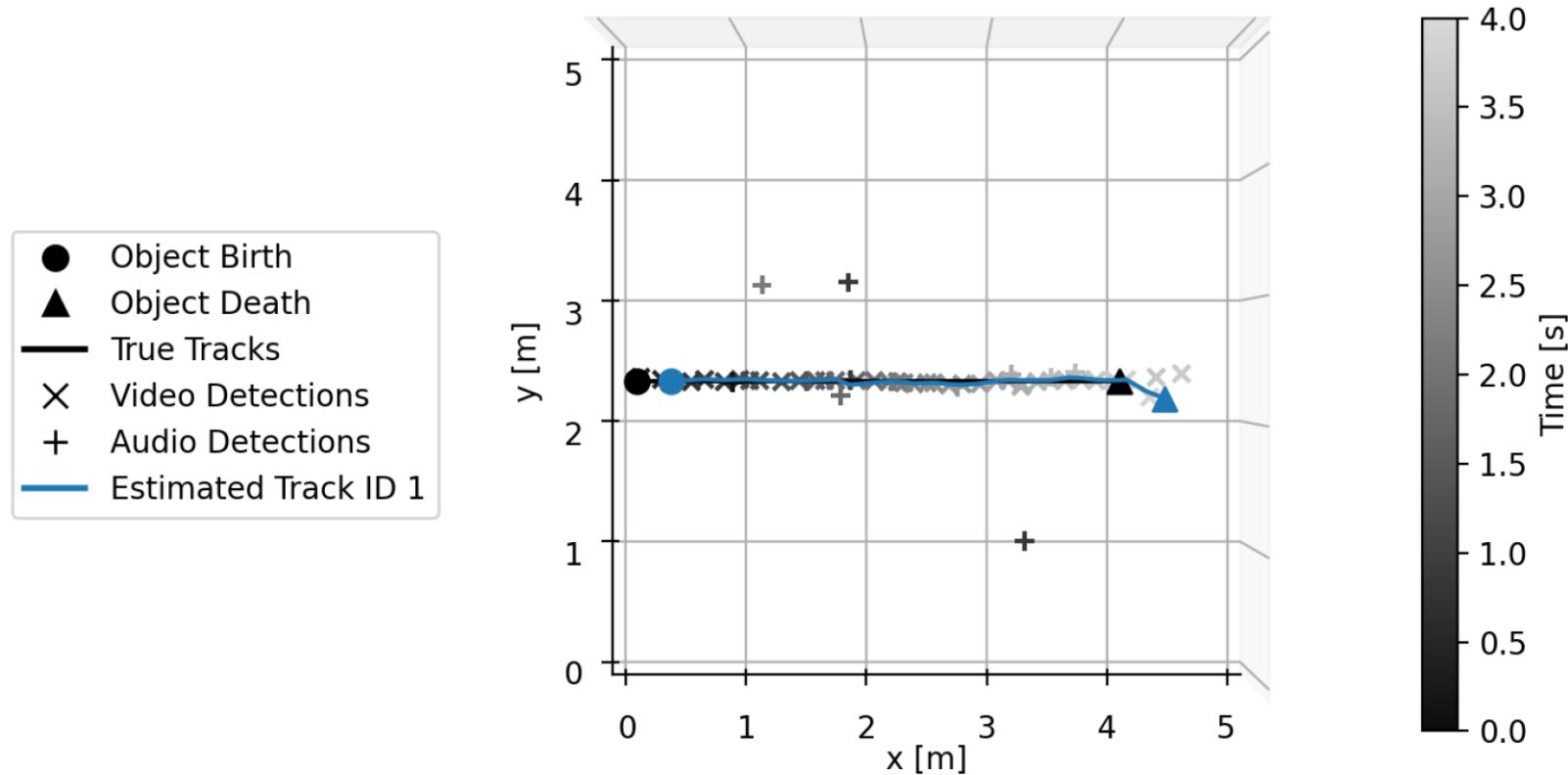
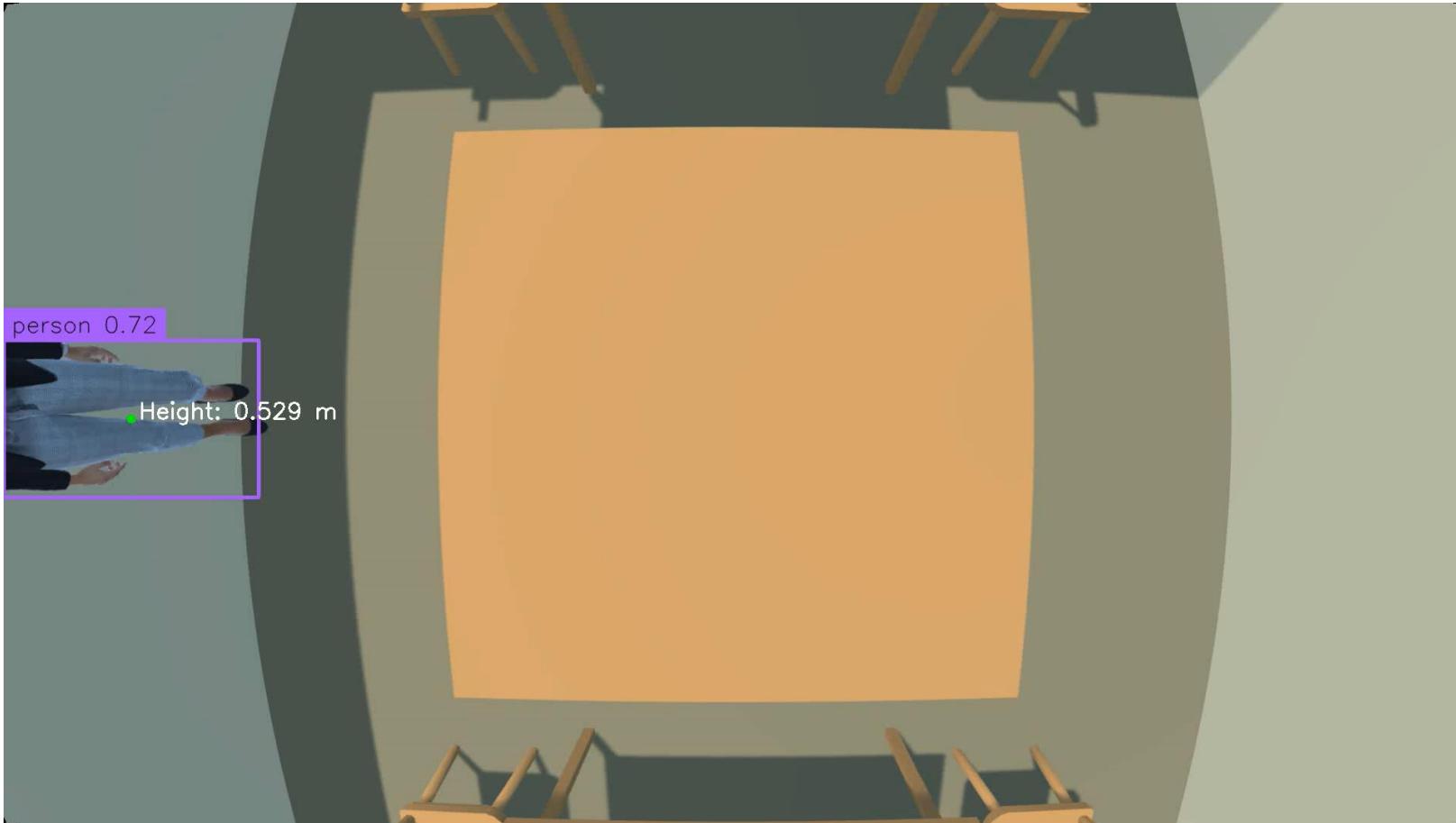


Figure 5.3: Tracking results of scenario S1 and modality M3.

Szenario S2: 1 Person läuft und spricht kontinuierlich, ist jedoch visuell temporär verdeckt



Szenario S2: 1 Person läuft und spricht kontinuierlich, ist jedoch visuell temporär verdeckt

Modalität M1: Audio

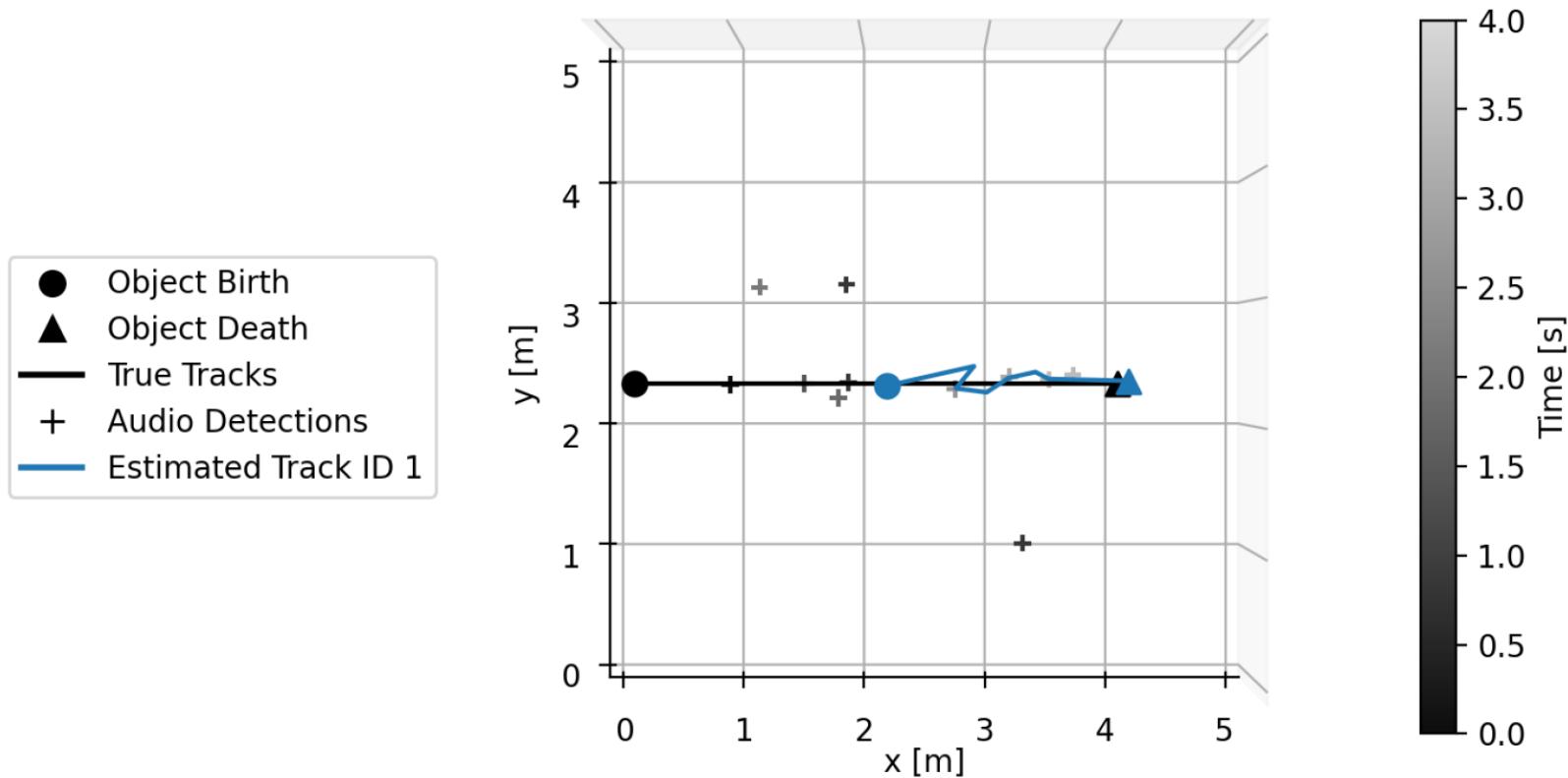


Figure 5.4: Tracking results of scenario S2 and modality M1.

Szenario S2: 1 Person läuft und spricht kontinuierlich, ist jedoch visuell temporär verdeckt

Modalität M2: Video

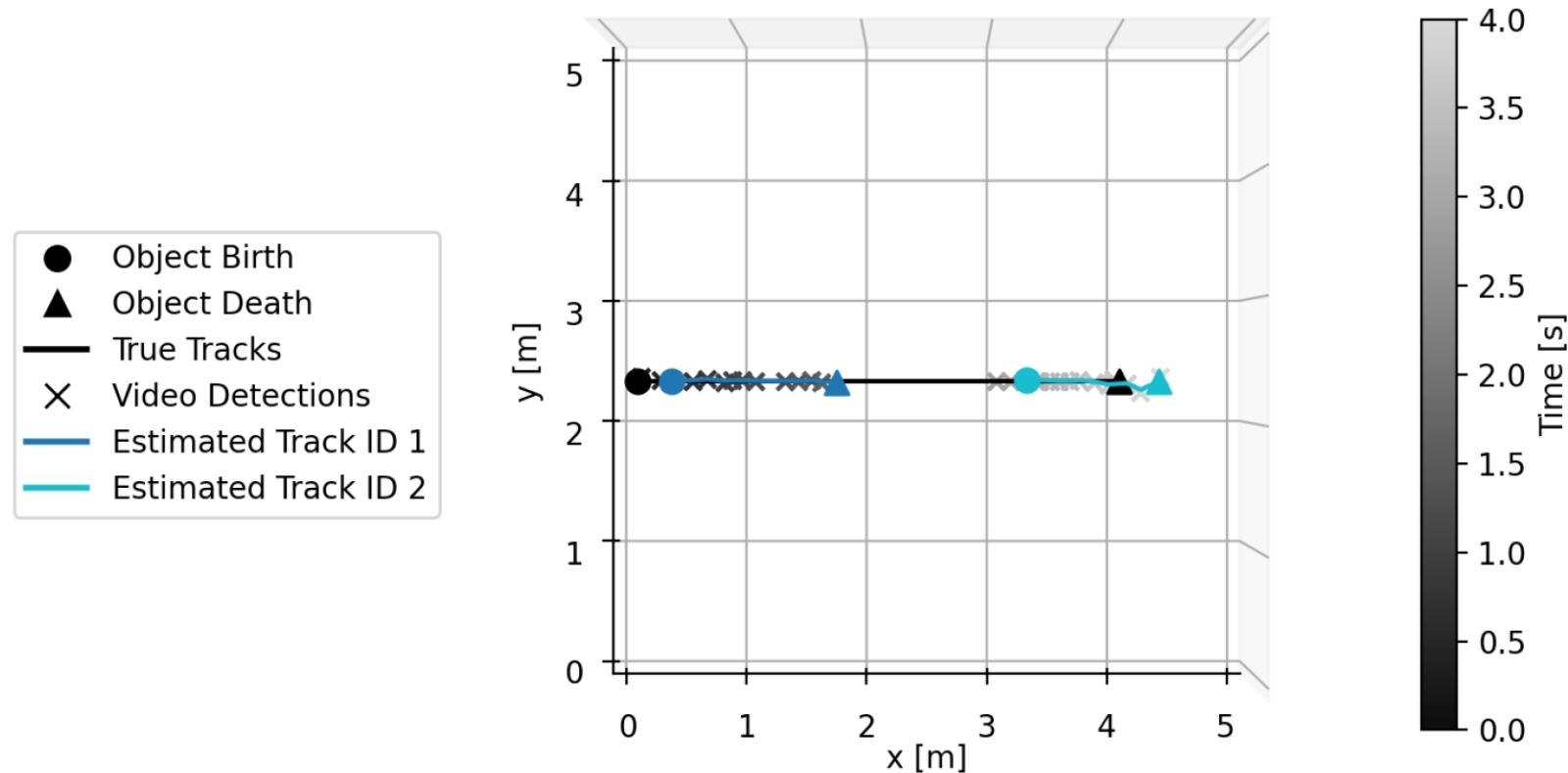


Figure 5.5: Tracking results of scenario S2 and modality M2.

Szenario S2: 1 Person läuft und spricht kontinuierlich, ist jedoch visuell temporär verdeckt

Modalität M3: Audio und Video

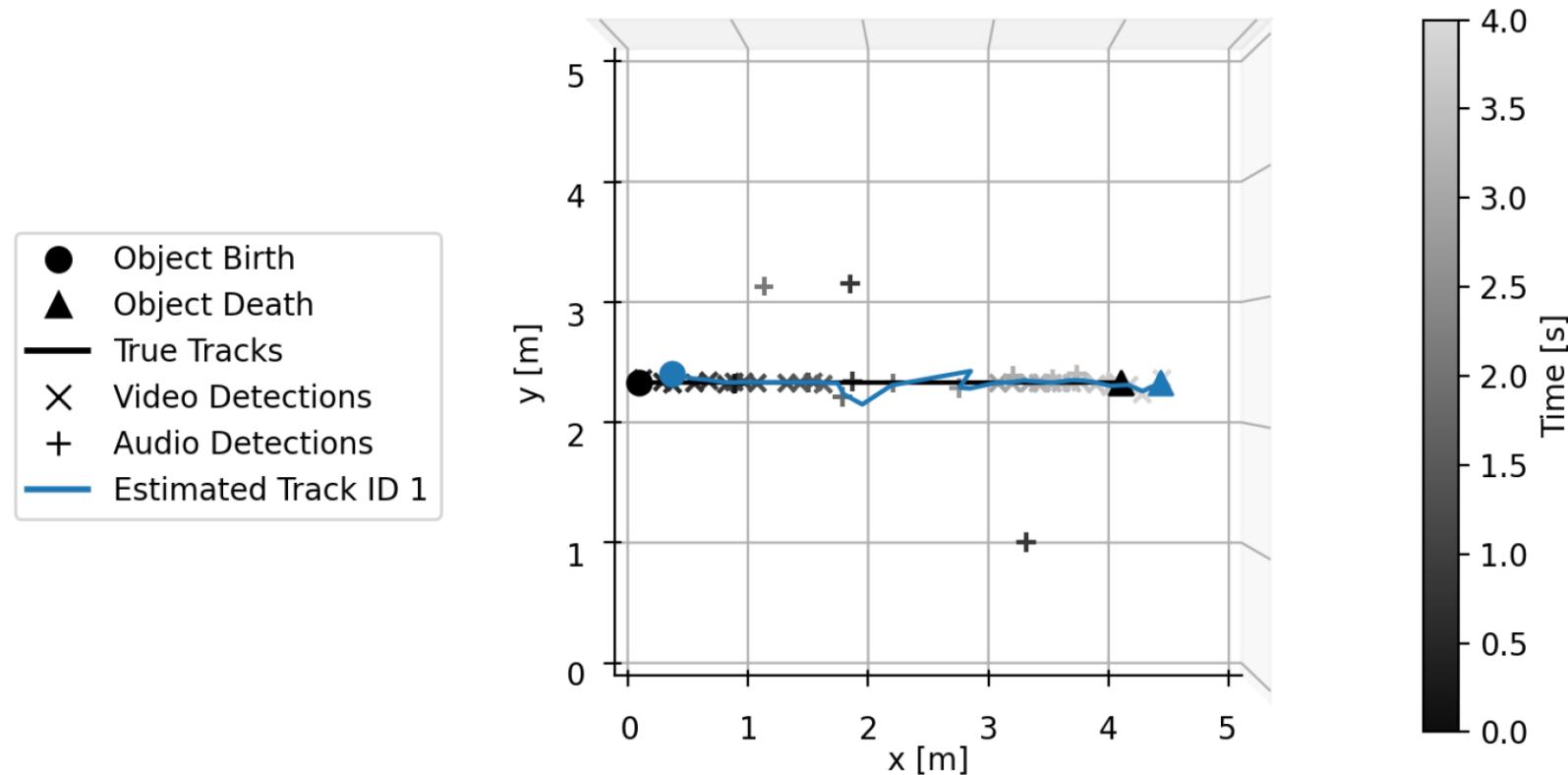


Figure 5.6: Tracking results of scenario S2 and modality M3.

Szenario S3: 3 Personen laufen. Die mittlere Person entgegengesetzt und spricht kontinuierlich



Szenario S3: 3 Personen laufen. Die mittlere Person entgegengesetzt und spricht kontinuierlich

Modalität M1: Audio

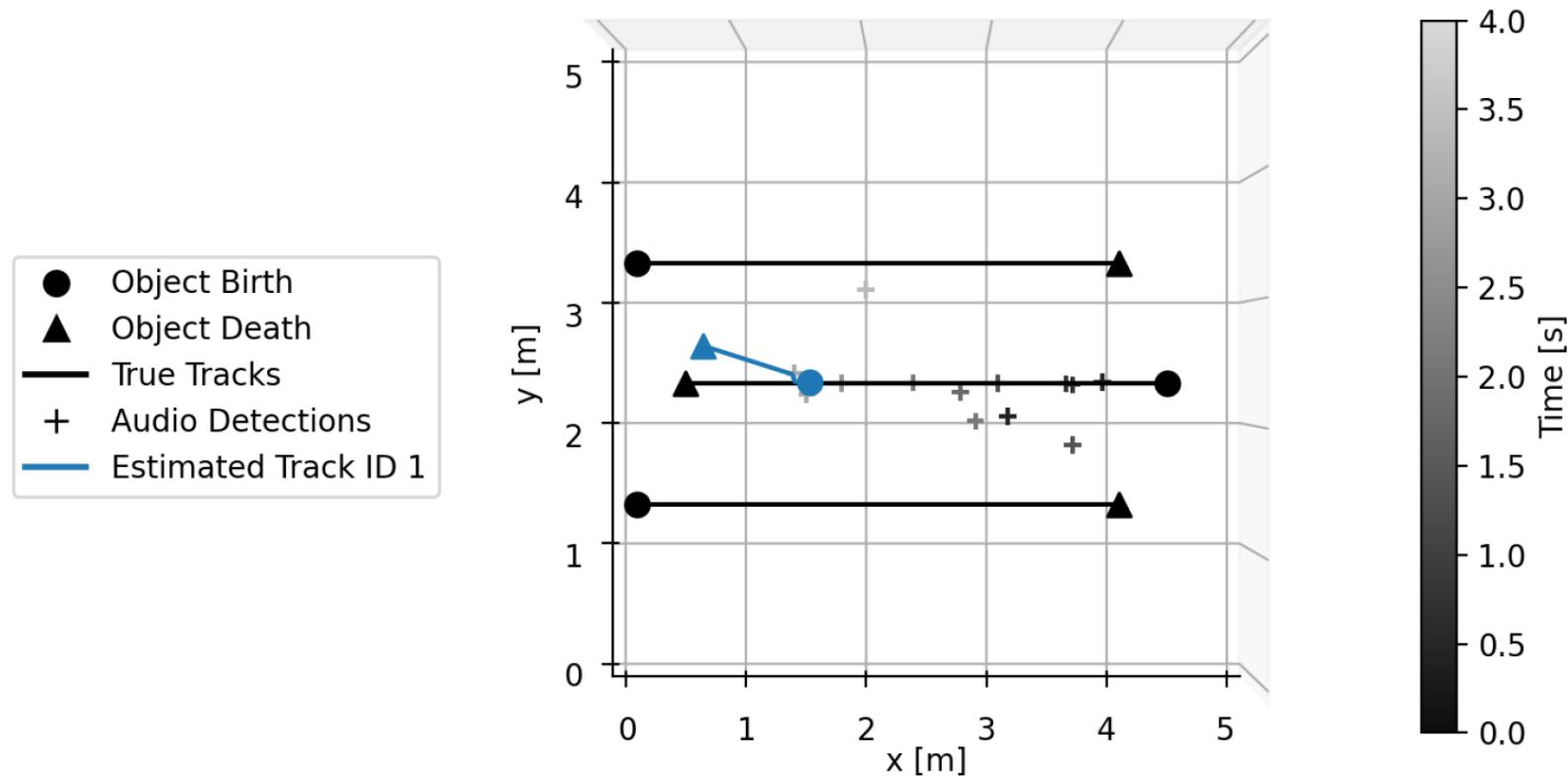


Figure 5.7: Tracking results of scenario S₃ and modality M₁.

Szenario S3: 3 Personen laufen. Die mittlere Person entgegengesetzt und spricht kontinuierlich

Modalität M2: Video

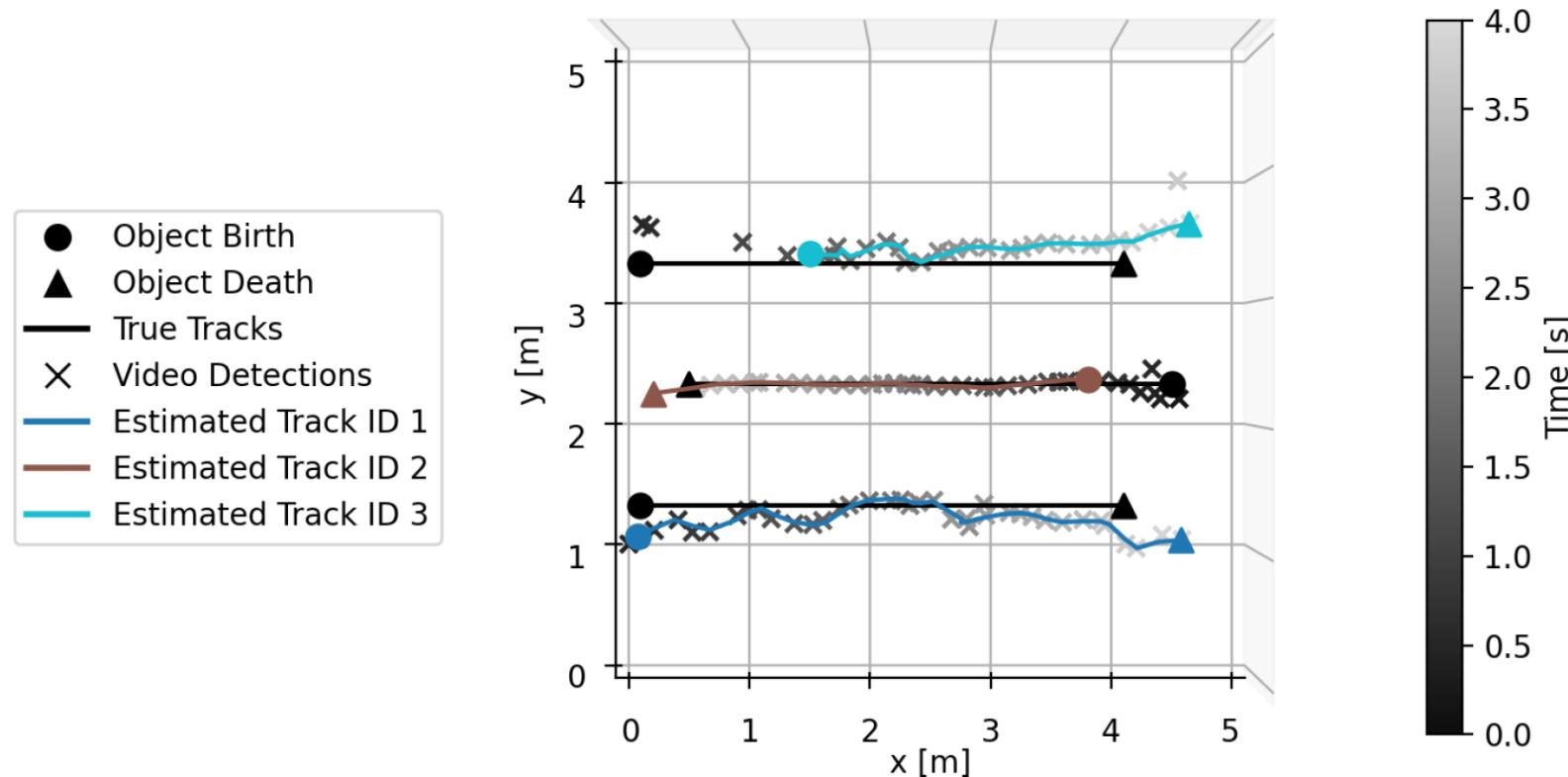


Figure 5.8: Tracking results of scenario S3 and modality M2.

Szenario S3: 3 Personen laufen. Die mittlere Person entgegengesetzt und spricht kontinuierlich

Modalität M3: Audio und Video

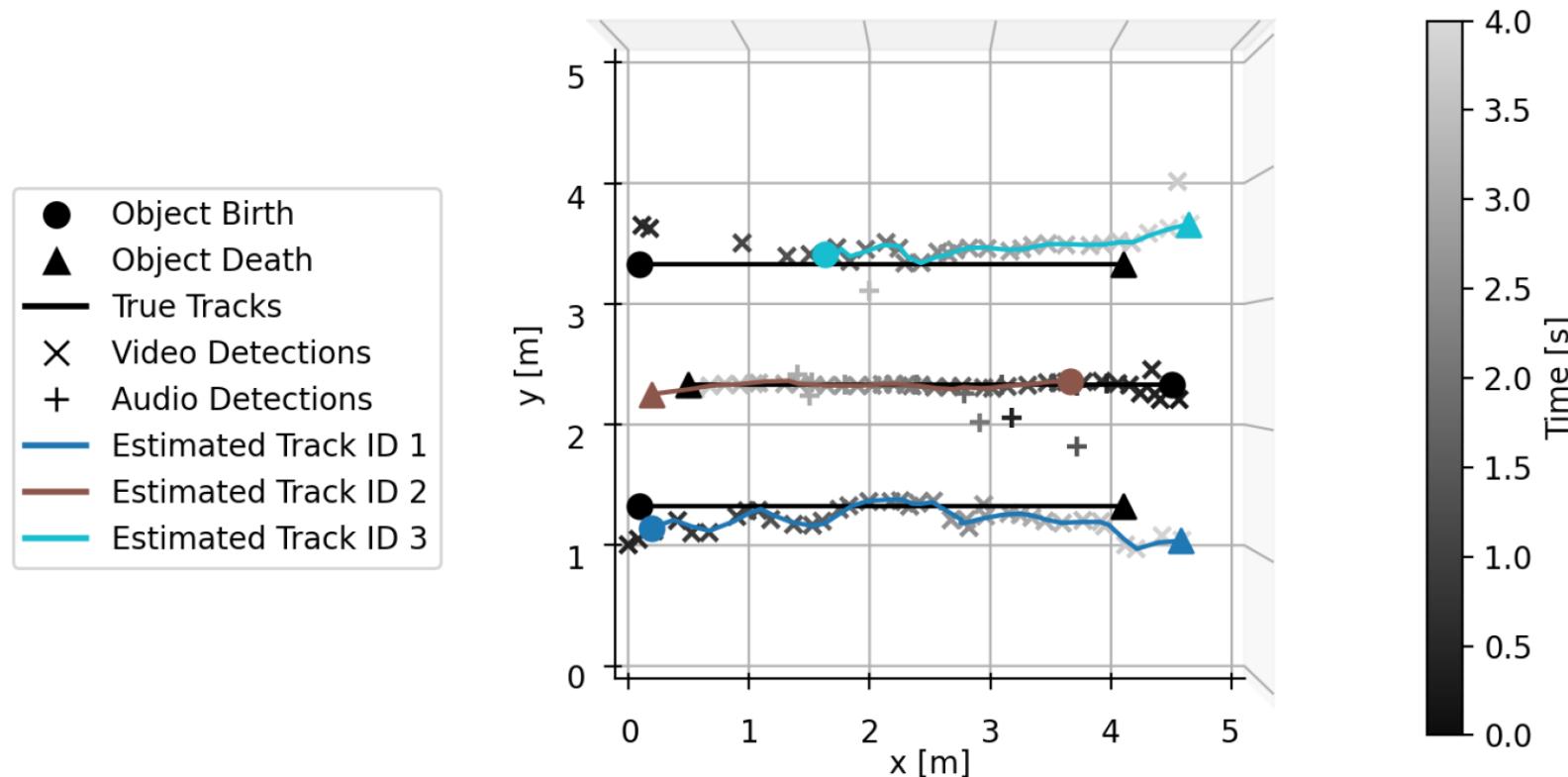


Figure 5.9: Tracking results of scenario S₃ and modality M₃.

Szenario S4: 3 Personen laufen. Die mittlere Person entgegengesetzt. Sie sprechen sequentiell 1/3 der Zeit



Szenario S4: 3 Personen laufen. Die mittlere Person entgegengesetzt. Sie sprechen sequentiell 1/3 der Zeit

Modalität M1: Audio

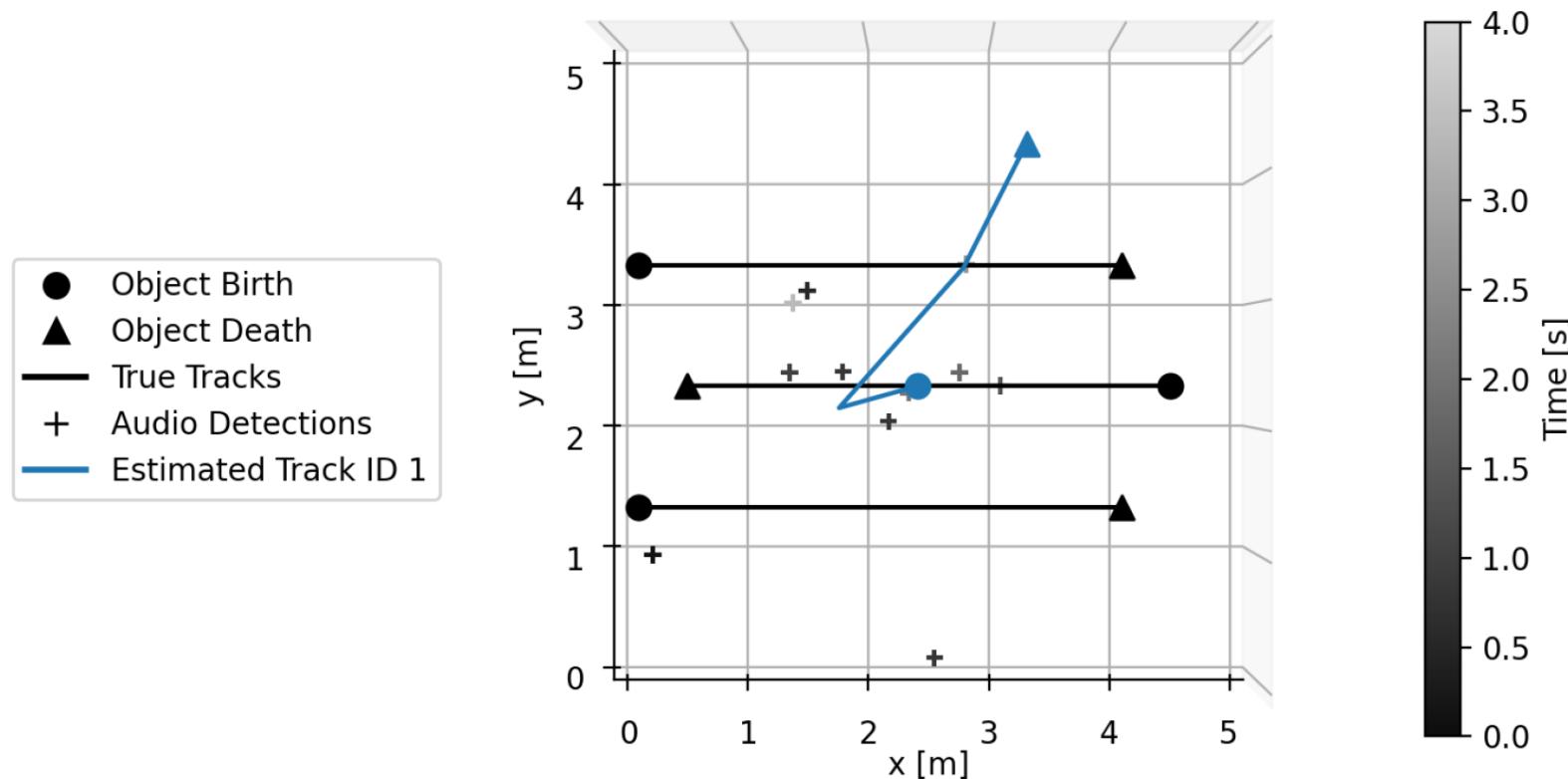


Figure 5.10: Tracking results of scenario S4 and modality M1.

Szenario S4: 3 Personen laufen. Die mittlere Person entgegengesetzt. Sie sprechen sequentiell 1/3 der Zeit

Modalität M2: Video

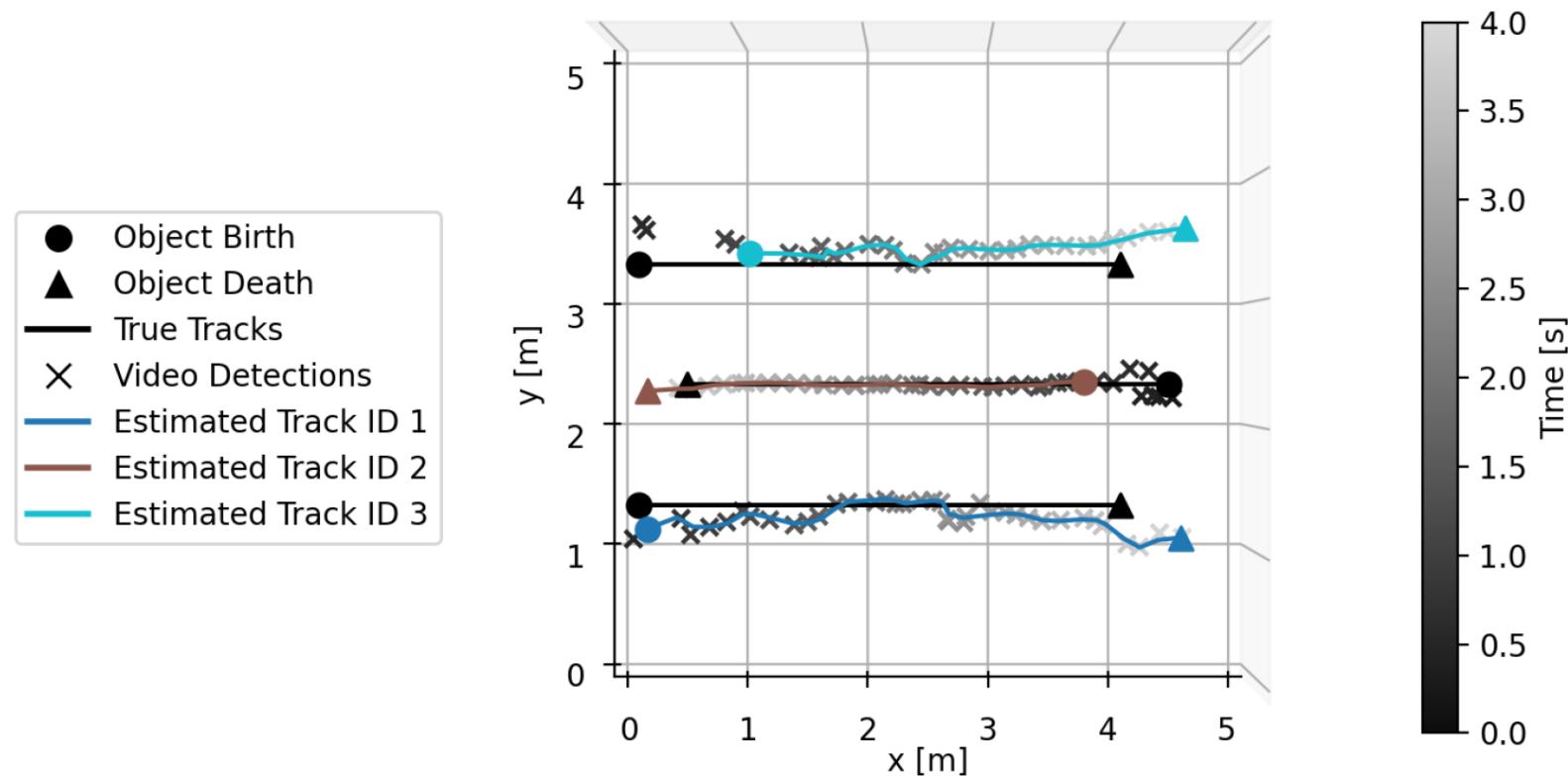


Figure 5.11: Tracking results of scenario S4 and modality M2.

Szenario S4: 3 Personen laufen. Die mittlere Person entgegengesetzt. Sie sprechen sequentiell 1/3 der Zeit

Modalität M3: Audio und Video

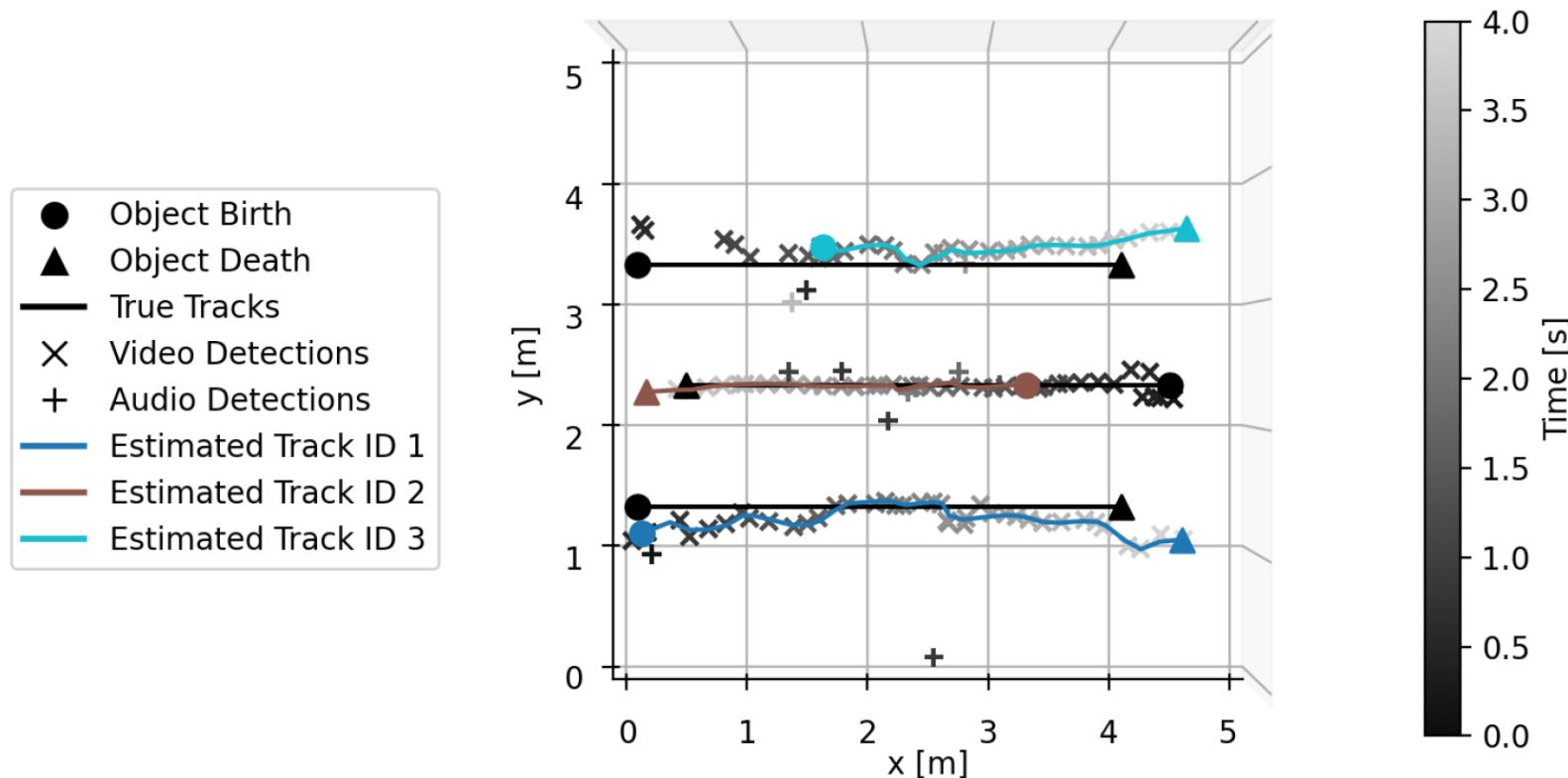


Figure 5.12: Tracking results of scenario S4 and modality M3.

Evaluation aller Szenarien

MOT25A (Audio M1)

- Hauptproblem: geringe Anzahl Detektionen, insbesondere in Sprechpausen
- Hohe Lokalisierungsgenauigkeit, wenn Detektionen vorhanden
- Audio alleine ungeeignet für Tracking von Sprechern

MOT25V (Video M2)

- Stabile Detektion & Assoziation
- Leistungseinbruch bei visueller Verdeckung
- Robustes Sprecher Tracking bei kontinuierlicher Sichtbarkeit

MOT25AV (Audio + Video M3)

- Kompensiert Modalitätsausfälle
- Größter Vorteil in Szenarien mit eingeschränktem Video
- In komplexen Szenen nicht immer besser als nur Video

Table 5.10: MS-GLMB tracking results of all scenarios

Modality	Combined Results Over All Scenarios							
	HOTA↑	DetA↑	AssA↑	DetRe↑	DetPr↑	AssRe↑	AssPr↑	LocA↑
M1 (Audio)	26.32	18.08	38.40	18.26	93.59	38.92	89.60	92.66
M2 (Video)	68.54	68.61	68.49	70.40	93.10	70.17	93.74	91.83
M3 (Audio+Video)	70.29	69.38	71.24	71.18	93.39	72.88	93.97	92.04

6. ZUSAMMENFASSUNG

Beantwortung der Forschungsfragen und Hypothesen

F1 - Simulation realistischer AV-Daten:

- Entwicklung einer vollständigen Audio- & Video-Simulationspipeline
- Realistische Raumakustik & geometrisch konsistente Bilddaten
- Erzeugung synthetischer Datensätze: MOT25A, MOT25V, MOT25AV

F2 - Feature-Level-Sensorfusion:

- Modulare Software Architektur mit Standard Schnittstellen
- Fokus auf Synchronisation, Austauschbarkeit & Effizienz
- Feature-Level-Sensorfusion (Tracking-by-Detection)

F3 / Evaluation – Tracking-Performance

- **MOT25A** bestätigt **H1**: Audio ist nicht robust bei Sprechpausen
- **MOT25V** bestätigt **H2**: Video ist leistungsstark, aber verdeckungsanfällig
- **MOT25AV** bestätigt **H3**: Audio-Video-Fusion erzielt den höchsten HOTA-Score

7. AUSBLICK

Ausblick

Erweiterung der Simulation

- Monokulare 3D-Lokalisierung ohne Tiefeninformation
- Kontrollierte Störfaktoren zur Robustheitsanalyse der Detektoren
- Anpassung von Sensorpositionen

Erweiterung der Datensätze

- Längere Szenarien mit vielen Personen
- Verschiedene Sprecherposen (z. B. Sitzen, Stehen)
- Räumliche Ausstattung

Methodische Erweiterungen

- Vergleich mit anderen State-of-the-Art Detektoren & MOT Algorithmen
- Algorithmen mit wenigen Parametern und unter Berücksichtigung von nicht-gleichförmigen Bewegungen

Bedeutung für Forschung und Praxis

- Annotierte Datengrundlage für Training & Validierung von:
 - Machine Learning Audio Detektoren & Multi-Object Trackern
- Vergleich synthetischer vs. realer Aufnahmen
- Deployment als reale Anwendungen (z. B. Gebäudeautomation)

8. OFFENE DISKUSSION & FRAGEN

Literatur- und Abbildungsverzeichnis

- [A] O. Rudolf, R. Hecker, M. Thisen, L. Sillekens, I. Penner, J.-P. Akelbein, S. Seyfarth, and Elke Hergenrother. “Implementation of visual people counting algorithms in embedded systems.” In: Computer Science Research
- [B] <https://www.sensortips.com/wp-content/uploads/2021/08/Different-fusion-levels-for-sensor-information.jpg>
- [C] <https://www.sensortips.com/wp-content/uploads/2021/08/Competitive-complementary-and-cooperative-sensor-fusion.jpg>
- [D] Xuan, J. O. S. (2021). Online Audio-Visual Multi-Source Tracking and Separation: A Labeled Random Finite Set Approach.
- [E] <https://de.mathworks.com/videos/understanding-kalman-filters-part-3-optimal-state-estimator--1490710645421.html>
- [F] Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., & Leibe, B. (2021). Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2), 548-578.

Synthetische Audiodatengenerierung

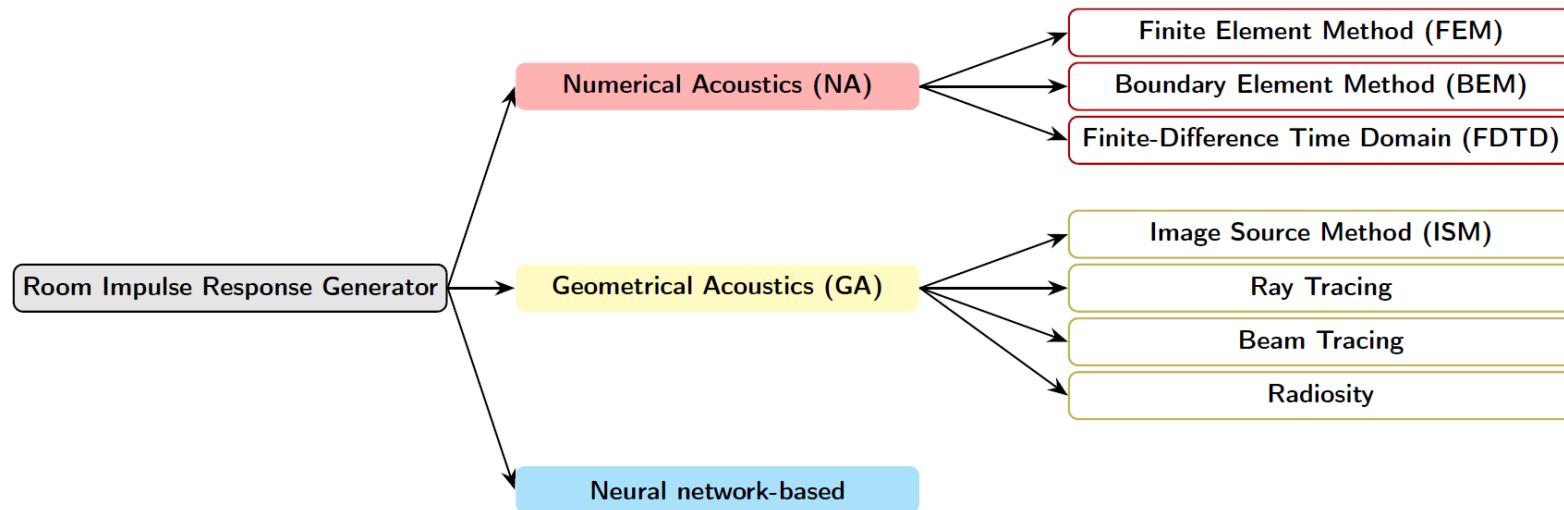


Figure 2.4: Classification of Room Impulse Response generator methods.

Deep-Learning-basierte Objekterkennung

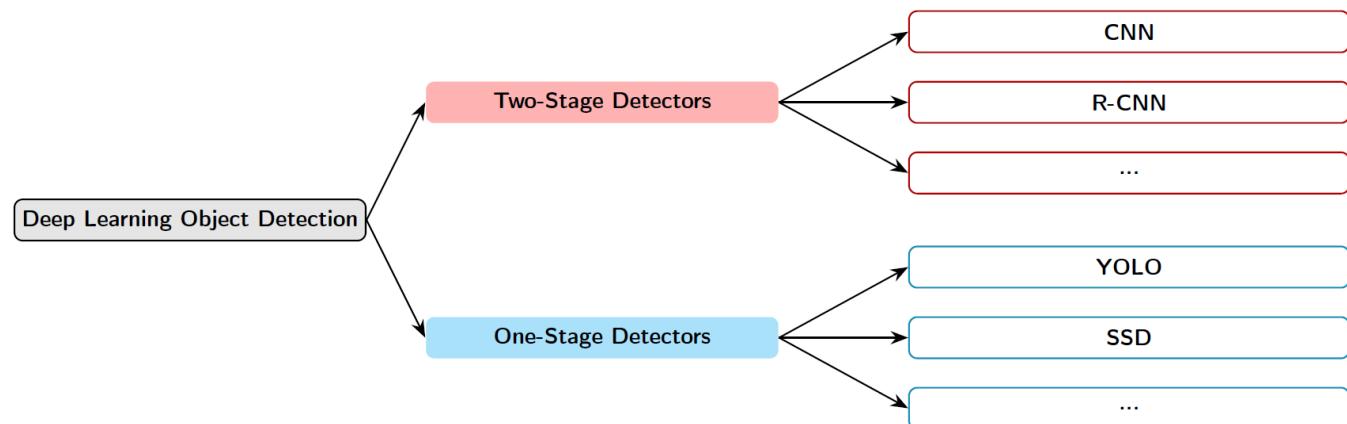


Figure 2.10: Classification of object detection methods.

Multi-object Tracking

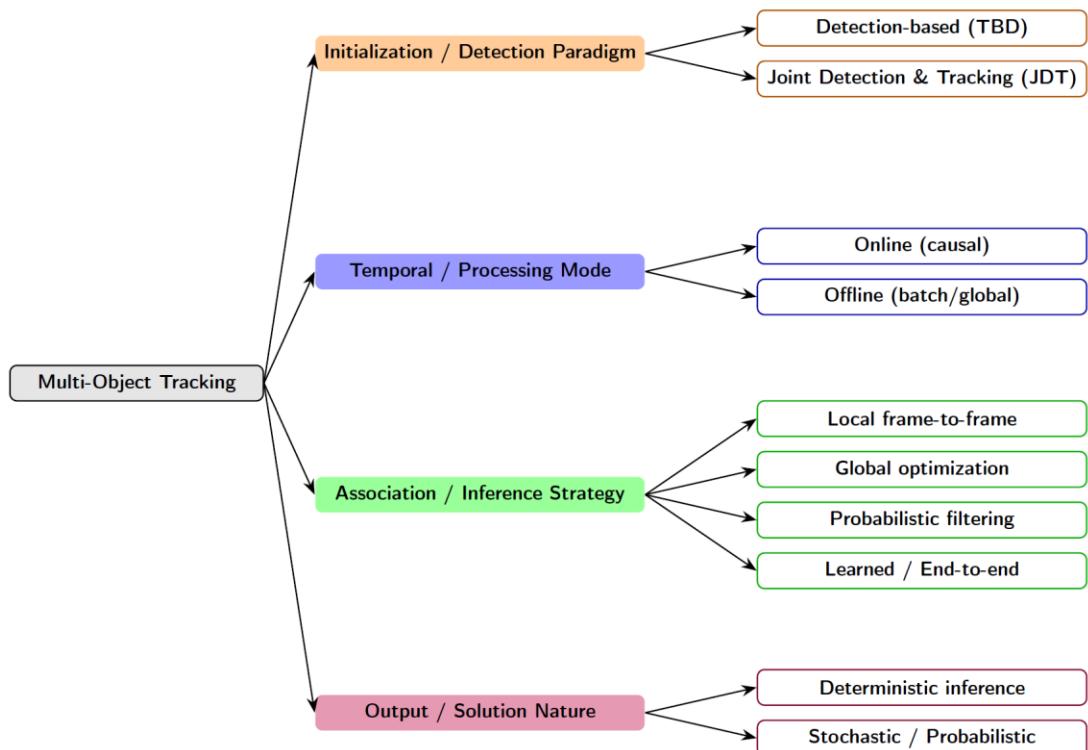


Figure 2.11: Classification perspectives for Multi-Object Tracking.

Table 2.4: Overview and comparison of multi-object tracking algorithms.

Algorithm	Detection Paradigm	Processing Mode	Inference Strategy	Solution Nature
SORT [79]	TBD	Online	Local	Deterministic
DeepSORT [80]	TBD	Online	Local	Deterministic
ByteTrack [81]	TBD	Online	Local	Deterministic
FairMOT [82]	JDT	Online	Learned	Deterministic
MOTR [83]	JDT	Offline	Learned	Deterministic
TrackFormer [84]	JDT	Offline	Learned	Deterministic
MS-GLMB [85]	TBD	Online	Probabilistic Filtering	Stochastic
HybridTrack [78]	JDT	Online	Local + Learned	Stochastic

Multi-Object Tracking

Multi-Sensor Generalized Labeled Multi-Bernoulli Filter (MS-GLMB)

Problemstellung und Ziel:

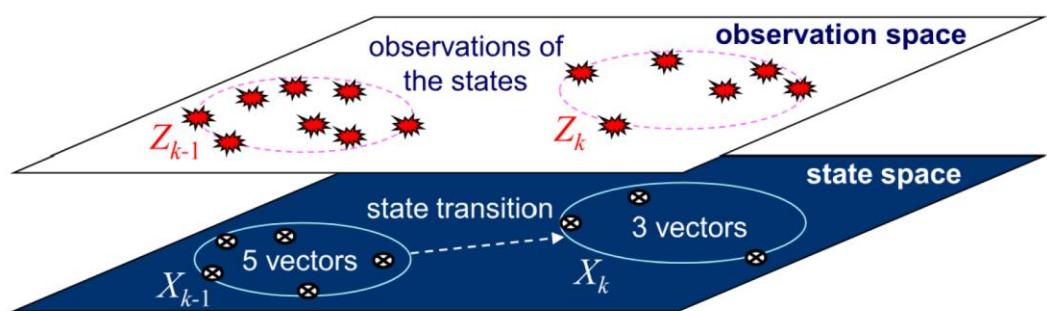


Figure 2.13: The State Space Model and Observation Space Model of a Multi-object system. [88].

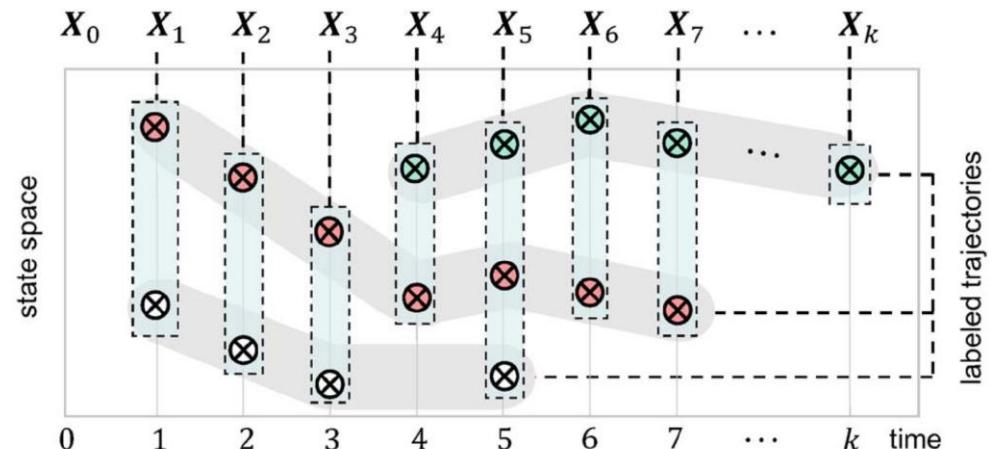
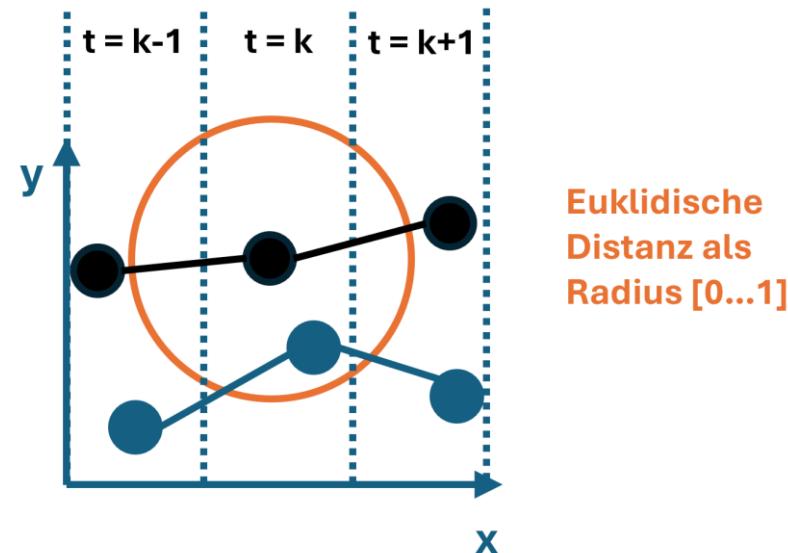


Figure 2.12: Labeled assignments in Random Finite Sets represent multi-object states and trajectories. [88]

Multi-Object Tracking Evaluation

Higher Order Tracking Accuracy (HOA)

Ähnlichkeitswert: Euklidische Distanz



$$S_{\text{Euclid}}(p_{pr}, p_{gt}) = \max \left[0, 1 - \frac{d(p_{pr}, p_{gt})}{d_0} \right] \quad (2.9)$$

HOA-Metrik und Sub-Metriken

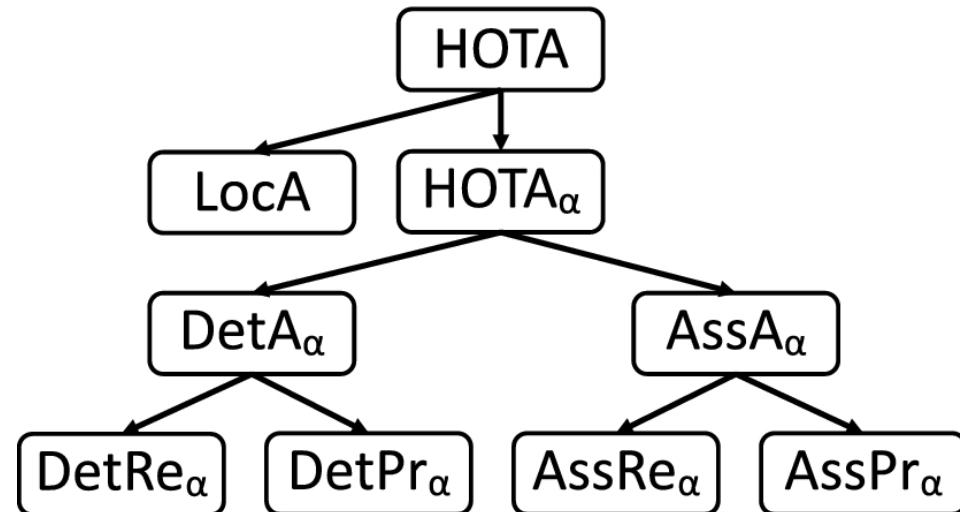


Figure F: HOA metric and its sub-metrics with values between 0 and 100.