VDI

REIHE 10
INFORMATIK/
KOMMUNIKATION

Fortschritt-Berichte VDI

Dipl.-Math. Roberto D. Henschel,
Hannover

NR. 875

Higher-Order Multiple Object Tracking

BAND
1|1

VOLUME
1|1

tnt
Institut für Informationsverarbeitung
www.tnt.uni-hannover.de

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Higher-Order Multiple Object Tracking

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des akademischen Grades
Doktor-Ingenieur
(abgekürzt: Dr.-Ing.)
genehmigte Dissertation

von

Dipl.-Math. Roberto D. Henschel
geboren am 10. September 1985 in Berlin, Deutschland

2021

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

1. Referent:	Prof. Dr.-Ing. Bodo Rosenhahn
2. Referent:	Prof. Dr.-Ing. habil. Christian Heipke
Vorsitzender:	Prof. Dr.-Ing. Jörn Ostermann

Tag der Promotion: 01. Oktober 2021

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

REIHE 10
INFORMATIK/
KOMMUNIKATION

Fortschritt-Berichte VDI

Dipl.-Math. Roberto D. Henschel,
Hannover

NR. 875

Higher-Order Multiple Object Tracking

BAND
1|1

VOLUME
1|1

Institut für Informationsverarbeitung
www.tnt.uni-hannover.de

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

VDI verlag

Henschel, Roberto D.
Higher-Order Multiple Object Tracking
Fortschritt-Berichte VDI, Reihe 10, Nr. 875. Düsseldorf: VDI Verlag 2021.
212 Seiten, 67 Bilder, 16 Tabellen.
ISBN 987-3-18-387510-7, ISSN 0178-9627,
76,00 EUR/VDI-Mitgliederpreis: 68,40

Für die Dokumentation: Verfolgung Mehrerer Objekte – Binäres Lineares Programm (BLP) – Binäres Quadratisches Programm (BQP) – Datenassoziationsmodel Höherer Ordnung – Globale Optimierung – Video – Sensorfusion – Inertiale Messeinheit

Keywords: Multiple Object Tracking – Binary Linear Programming (BLP) – Binary Quadratic Programming (BQP) – Higher-Order Data Association Model – Global Optimization – Video – Sensor Fusion – Inertial Sensors

This dissertation deals with camera-based offline multiple object tracking and explores higher-order data association models. Due to their extensive exploitation of the available information, such models are promising approaches in current research. However, they commonly represent NP-hard optimization problems so that their application in practice is challenging.
The first part of this thesis proposes a binary quadratic program that enables to globally fuse signals within a higher-order data association model. This enables to overcome weaknesses of the individual signals. An approximate solver based on the Frank-Wolfe algorithm is presented and analyzed. Its benefit is demonstrated in two setups: fusion of two detectors and combining signals coming from a video and body-worn inertial measurement units. The second part of this thesis proposes an extension of the disjoint path model by higher-order information and connectivity priors, resulting in a binary linear program. Efficient separation algorithms are proposed and integrated into a cutting-plane algorithm, making it possible for the first time to solve higher-order data association globally in practice.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek (German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at www.dnb.de.

© VDI Verlag GmbH | Düsseldorf 2021
Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe (Fotokopie, Mikrokopie), der Speicherung in Datenverarbeitungsanlagen, im Internet und das der Übersetzung, vorbehalten. Als Manuskript gedruckt. Printed in Germany.

ISBN 987-3-18-387510-7, ISSN 0178-9627

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Acknowledgments

This thesis was written in the course of my activity as a scientific research assistant at the *Institut für Informationsverarbeitung* of the Leibniz University Hannover.

First of all, I would like to thank my doctoral advisor Prof. Dr.-Ing. Bodo Rosenhahn for supporting me to become a researcher in the very exciting field of computer vision. I am thankful for the stimulating discussions about work, society, and mathematical topics, the great supervision, and all the support and freedom I received to pursue my ideas.

Also, many thanks to him and Prof. Dr.-Ing. Jörn Ostermann for the excellent working environment. I also thank Prof. Dr.-Ing. habil. Christian Heipke for being the second examiner and Prof. Dr.-Ing. Jörn Ostermann for being the chair of the defense committee.

Special thanks go to Prof. Dr.-Ing. Laura Leal-Taixé for her incredible support and deep discussions, which greatly helped to advance my research and this thesis.

Also, I would like to thank Andrea Hornakova and Dr. Paul Swoboda for the great collaboration, which was very inspiring, successful, and fun.

I would also like to thank Prof. Dr. Konrad Schindler for the opportunity to stay at his research lab, which was really inspiring.

The time at the TNT institute was amazing thanks to all the great colleagues. My special recognition goes to my former office mate Dr.-Ing. Timo von Marcard for making our office such a great place. Having deep discussions about research, mathematics, and non-work topics was great. Also, collaborating and traveling together was outstanding as was all the support I received during the time when I was finishing my thesis. I would like to thank my office mate Timo Kaiser for the excellent time with lots of discussions about anything and for the fun working together on multiple object tracking. Also, I thank Dr.-Ing. Bastian Wandt for many deep discussions and for the joint building of a company. Special thanks go to my colleagues Dr.-Ing. Michele Fenzi, Leonid German, Dr.-Ing. Alina Kutznetsova, Prof. Dr.-Ing. Laura Leal-Taixé, Yasser Samayoa, and Dr.-Ing. Dipl.-Math. techn. Aron Sommer for the great fun at the institute, *e.g.*, table football tournaments, card games or long discussions, and all the fun inside or outside the institute, making the time together unforgettable. Finally, I would like to thank the TNT administrative staff for their technical and administrative support.

I thank my friend Gad Kohls for advising me to study mathematics.

Last but not least, I dedicate a special thanks to my family. Without the unconditional support from my parents Clara and Ramon, as well as my sister Ruth, accomplishing my degrees and this thesis would not have been possible.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Contents

1 Introduction **1**

1.1 Applications . . . 1

1.2 The Multiple Object Tracking Problem . . . 3

1.2.1 Video-based *Multiple Object Tracking* (MOT) . . . 4

1.2.2 Tracking-by-detection . . . 7

1.3 Challenges of Multiple Object Tracking . . . 11

1.3.1 Errors caused by the object detector . . . 12

1.3.2 Challenges in discriminative features . . . 13

1.3.3 Combinatorial challenges . . . 16

1.4 Related Work . . . 18

1.5 Contributions . . . 21

1.6 List of Publications . . . 26

1.7 Outline . . . 30

2 Fundamentals **33**

2.1 Sets, Maps, and Matrices . . . 33

2.2 Probability Theory . . . 34

2.3 Graph Theory . . . 34

2.3.1 Important graph classes . . . 36

2.3.2 Computations on graphs . . . 37

2.4 Machine Learning . . . 38

2.4.1 Supervised learning . . . 39

2.4.2 Logistic regression . . . 40

2.4.3 Neural networks . . . 41

2.5 Computational Complexity Theory . . . 47

2.6 Optimization Theory . . . 52

2.6.1 Linear programming . . . 52

2.6.2 Binary linear programming . . . 54

2.6.3 Quadratic programming . . . 59

2.6.4 Binary quadratic programming . . . 60

2.6.5 Non-linear optimization . . . 61

2.7 Multi-Object Tracking . . . 61

2.7.1 Object detectors . . . 61

2.7.2 Appearance features . . . 66

2.7.3 Datasets . . . 75

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

2.7.4 MOT metrics . . . 76

3 HO-MOT with Signal Fusion 80
3.1 Introduction . . . 81
3.2 Signal Fusion as Weighted Graph Labeling Problem . . . 83
3.2.1 Related work . . . 84
3.2.2 Data association model for signal fusion . . . 84
3.3 Frank-Wolfe Optimizer for Weighted Graph Labeling Problems . . . 89
3.3.1 Related work . . . 90
3.3.2 Frank-Wolfe Optimizer for Binary Solutions . . . 91
3.4 Multiple People Tracking by Fusing Head and People Detections . . . 98
3.4.1 Data association model . . . 99
3.4.2 Experimental results . . . 102
3.5 Simultaneous Identification and Tracking of Multiple People using Video and IMUs . . . 106
3.5.1 Related work . . . 110
3.5.2 Method . . . 112
3.5.3 Evaluation . . . 117
3.6 Conclusion . . . 127

4 Lifted Disjoint Paths 130
4.1 Introduction . . . 131
4.2 Related Work . . . 132
4.3 Problem Formulation . . . 134
4.4 Constraints . . . 135
4.5 Separation . . . 144
4.6 Complexity . . . 146
4.7 Experiments . . . 150
4.7.1 Graph construction. . . . 150
4.7.2 Pre-processing and post-processing . . . 151
4.7.3 Cost learning . . . 152
4.7.4 Implementation details on the lifted disjoint paths solver . . . 156
4.7.5 Experiment setup . . . 156
4.7.6 Benefit of long-range edges . . . 157
4.7.7 Ablation study on post-processing methods. . . . 157
4.7.8 Accuracy of the fusion network . . . 159
4.7.9 Qualitative evaluations . . . 159
4.7.10 Benchmark evaluations . . . 160
4.8 Conclusion . . . 161

5 Conclusions 165

Bibliography 172

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Acronyms

2D	*two-dimensional*
3D	*three-dimensional*
ACF	*Aggregate Channel Features*
Acc	*Accuracy*
BCE	*Binary Cross-Entropy*
BLP	*Binary Linear Program*
BQP	*Binary Quadratic Program*
CE	*Cross Entropy*
CNN	*Convolutional Neural Network*
CS	*Cosine Similarity*
DM	*DeepMatching*
DP	*Disjoint Paths*
DPM	*Deformable Part Model*
FN	*False Negatives*
FP	*False Positives*
fps	*frames per second*
FRCNN	*Faster R-CNN*
FW	*Frank-Wolfe*
FWT	*Frank-Wolfe Tracker*
GAP	*Duality gap*
GPS	*Global Positioning System*

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

HO-MOT	*Higher-Order Multiple Object Tracking*
HOG	*Histogram of Oriented Gradients*
ID	*Identity*
IDF1	*ID F1*
IDP	*ID Precision*
IDR	*ID Recall*
IDS	*ID Switches*
IMU	*Inertial Measurement Unit*
IoU	*Intersection over Union*
KLT	*Kanade–Lucas–Tomasi feature tracker*
LDP	*Lifted Disjoint Paths*
LIDAR	*Light Detection and Ranging*
Lif_T	*Lifted Disjoint Paths Tracker*
Lif_TsimInt	*Lifted Disjoint Paths Tracker using simple linear interpolation*
LP	*Linear Program*
mAP	*mean Average Precision*
ML	*Mostly Lost*
MOT	*Multiple Object Tracking*
MOTA	*Multiple Object Tracking Accuracy*
MPT	*Multiple People Tracking*
MSE	*Mean Squared Error*
MT	*Mostly Tracked*
PC	*Perspective Correction*
Prec	*Precision*
PT	*Partially Tracked*
QP	*Quadratic Program*
RADAR	*Radio Detection And Ranging*
ReLU	*Rectified Linear Unit*
RPN	*Region Proposal Network*

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

SDP	*Scale-Dependent Pooling*
SVM	*Support Vector Machine*
TPR	*True Positive Rate*
TNR	*True Negative Rate*
VHN	*Visual Heading Network*
VIMPT	*Video Inertial Multiple People Tracking*
VIT	*Video Inertial Tracker*

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Notation

Numbers and Arrays

a	A scalar
A	A set
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
$\mathbf{A}^\intercal$	Transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of square matrix $\mathbf{A}$
$\langle \mathbf{a}, \mathbf{b} \rangle$	Scalar product of $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{a} \star \mathbf{b}$	Convolution of $\mathbf{a}$ and $\mathbf{b}$
$\frac{df}{dx}$	Derivative of f with respect to x
$\frac{d^2 f}{dx^2}$	Second derivative of f with respect to x
∇f	Gradient of f
$\lfloor x \rfloor$	Integer part of x
$\|\|\mathbf{a}\|\|$	L^2-norm of $\mathbf{a}$
$\det(\mathbf{A})$	Determinant of $\mathbf{A}$
$\mathbf{I}$	Identity matrix
$\mathbf{1}$	Matrix of ones
$\mathbf{0}$	Zero matrix
$\mathbb{E}[X]$	Expectation of random variable X
$[n]$	Set of natural numbers from 1 to n
$[n]_0$	Set of natural numbers from 0 to n
$[n_1 : n_2]$	Set of natural numbers from n_1 to n_2

Symbols

$\mathfrak{f}$	Frame index
n_{R}	Number of frames in a recording
R	Set of frames indices of a recording

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

$\mathbf{d} \in D$	A detection $\mathbf{d}$ in a set of detections D
$D_{\mathfrak{f}}$	Set of all detections in frame $\mathfrak{f}$
$\gamma \in \Gamma$	A trajectory γ in a set of trajectories Γ
$\mathrm{supp}(\gamma)$	Set of frames for which trajectory γ contains detections
ϕ	Unary feature
ψ	Pairwise feature
$\mathcal{P}$	Deterministic polynomial time complexity class
$\mathcal{NP}$	Nondeterministic polynomial time complexity class
L	A loss
$\wedge$	Logical AND
$\vee$	Logical OR
$\mathcal{O}(n)$	Big O notation
P	Probability measure

Graphs

$\mathcal{G}$	A graph
$v \in \mathcal{V}$	A vertex v in a vertex set $\mathcal{V}$
$e \in \mathcal{E}$	An edge e in an edge set $\mathcal{E}$
$\mathcal{G}_{\mathcal{V}}$	The vertex set of graph $\mathcal{G}$
$\mathcal{G}_{\mathcal{E}}$	The edge set of graph $\mathcal{G}$
n_{nod}	Number of nodes
P	A path
$vw\text{-paths}(\mathcal{G})$	The set of paths in $\mathcal{G}$ starting at v and ending in w
$\mathcal{G}[\widetilde{\mathcal{V}}]$	The subgraph of $\mathcal{G}$ induced by the vertex set $\widetilde{\mathcal{V}}$
$d_G(v)$	Neighborhood of node v within the graph $\mathcal{G}$
$\mathcal{V}_{\mathfrak{f}}$	All nodes at frame $\mathfrak{f}$
$\mathbf{c}$	Vertex weights
$\mathbf{q}$	Edge weights
$\breve{\mathcal{G}} = (\breve{\mathcal{V}}, \breve{\mathcal{E}})$	Lifted graph $\breve{\mathcal{G}}$ with edge set $\breve{\mathcal{E}}$ and vertex set $\breve{\mathcal{V}}$
$\breve{q}_{\breve{e}}$	Weight of lifted edge $\breve{e}$
$\mathcal{R}$	Reachability relation
n_{obj}	Number of labels

Optimization

P	Polyhedron
$\mathrm{P}_{\mathcal{B}}$	Polyhedron P with additional binary constraints

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

$\mathrm{P}_{\mathcal{B}}^{\circ}$	Continuous relaxation of $\mathrm{P}_{\mathcal{B}}$
$\mathrm{P}(\mathbf{A},\mathbf{b})$	A polyhedron in canonical $\mathcal{H}$-representation
$\mathrm{conv}(A)$	Convex hull of a set A
$H^{\leq}$	Closed half-space
$\mathbf{x},\mathbf{y}$	Binary indicator variables
$[v \rightharpoonup k]$	Linear index to indicator variable for the assignment of node v to label k
$\mathrm{WGL}(\mathcal{G})$	Weighted graph labeling problem defined on graph $\mathcal{G}$
$\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$	Problem $\mathrm{WGL}(\mathcal{G})$ with additional non-maxima suppression
$\mathrm{RWGL}(\mathcal{G})$	Continuous relaxation of $\mathrm{WGL}(\mathcal{G})$
$\mathrm{P}_{\mathcal{B}}(\mathcal{G})$	Underlying polyhedron of problem $\mathrm{WGL}(\mathcal{G})$
$\mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G})$	Underlying polyhedron of problem $\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Abstract

This dissertation deals with methods for camera-based multiple object tracking (MOT). More precisely, the task is to compute the association between objects of a specified class and corresponding image contents of a video recording. To tackle this extremely difficult problem, the so-called tracking-by-detection paradigm is usually employed: First, object detections are generated for the entire recording. Then, an association between detections and objects is computed. Finding the correct assignment is called the data association problem. A solution to the problem provides the trajectories for the desired objects.

Since the tracking-by-detection paradigm is computed sequentially, errors of the detector as well as wrongly assessed temporal consistencies between detections can lead to propagation of errors. Therefore, the employed data association model substantially determines the accuracy of the computed trajectories. In order to achieve highly accurate results, this thesis focuses on building robust data association models. At the same time, standard detectors are used to meaningfully compare the corresponding results with previous approaches.

Many established methods use data association models that exploit temporal consistency only between detections that directly follow each other in a trajectory. However, such simple models are highly susceptible to the errors mentioned above. This thesis presents more robust tracking methods by using higher-order data association models (higher-order multiple object tracking). Here, all pairs of detections associated with a trajectory, and not only the consecutive ones, contribute to the evaluation of the consistency of a trajectory. To comprehensively exploit the entire information of a video recording, this thesis formulates two data association models, each as a global optimization problem. Accordingly, the recordings are processed offline, using all information available. However, the underlying optimization problems are $\mathcal{NP}$-hard, which makes it difficult to compute good solutions. A suitable optimization method is presented for each proposed data association model. The corresponding optimizer yields in practice near-optimal or even (to the best of our knowledge for the first time) provable global optimal solutions, depending on the model used.

In the first part of this thesis, a method for improved utilization of the signals available at a point in time is presented. While the standard tracking approach uses only object detections, the proposed method allows multiple input signals to be fused globally for the tracking task. A higher-order data association model is proposed that evaluates consistency within a signal as well as between different signals. By using complementary signals, weaknesses of individual signals can be compensated and advantages can be combined. The proposed data association model is based on an $\mathcal{NP}$-hard weighted graph labeling problem. Due to the complexity of the problem, computing an optimal solution is difficult. A suitable approximate optimization method for the graph labeling problem is presented. Evaluations show that near-optimal solutions are generated with this method. The benefits of the fusion are analyzed using two applications of the graph-labeling formulation. (i) To exploit more image information, person detections are combined with head detections. It is shown that the fusion achieves much better results than when only using person detections. In particular, the fusion helps to detect and remove false-positive person detections, as these often do not have matching head

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

detections. In addition, the fusion approach results in persons being tracked for longer periods of time, since in the case of missing person detections, head detections can be used to locate and track persons. (ii) A video is fused with inertial measurement units (IMU). For this purpose, it is assumed that each person to be tracked wears an IMU on his or her back. Acceleration and orientation measurements from the IMUs are linked with corresponding values estimated from the video recording. The corresponding graph labeling problem generates trajectories that are temporally consistent with respect to the video recording and the IMU signals. The fusion leads to significantly better trajectory results compared to purely video-based MOT methods, especially when the visual information is impaired (*e.g.*, due to motion blur or similarly dressed persons). Missing detections can be reconstructed very robustly by the fusion so that the method has a lower dependence on the quality of the detections compared to purely video-based methods. In addition, the proposed fusion allows people to be identified in the image since each trajectory is associated with an IMU. Overall, the methods from the first part of this thesis demonstrate that the proposed fusion formulation enables to exploit provided data more extensively by being able to process more image information and integrate more signals. This substantially improves tracking accuracy.

Nonetheless, object detections provide valuable information not fully exploited by existing methods. Higher-order data association models are either not used at all due to their complexity or are based on heuristic optimization methods. Both cases can lead to false associations.

In contrast, in the second part of this thesis, an optimization method is presented that allows for the first time to solve a suitable higher-order data association model by means of global optimization despite being $\mathcal{NP}$-hard. This enables to exploit long-term temporal information and long-range temporal interactions. To this end, a novel data association model is proposed and described as a binary linear program. Efficient separation algorithms are presented to solve the optimization problem within a cutting-plane method. The global optimization enabled the method to outperform the state of the art on all tested datasets by a large margin. In addition, conducted experiments show that the method benefits significantly from the use of long-term information. On the datasets used, the presented method leads to nearly optimal assignment accuracies for given detections. Future work can therefore focus on other areas, such as a more accurate extraction of detections. Overall, the second part of this thesis shows that improved exploitation of the information provided over time leads to substantial improvements in trajectory results. For the first time, a global optimization method has been successfully used to solve higher-order data association models.

Keywords: Multiple Object Tracking, Video, Higher-Order Data Association Models, Sensor Fusion, Binary Linear Program, Binary Quadratic Program, Global Optimal Solution

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Kurzfassung

Diese Dissertation befasst sich mit Verfahren zur kamerabasierten Verfolgung mehrerer Objekte (Multiple Object Tracking, abgekürzt MOT). Genauer besteht die Aufgabe in der Zuordnungsberechnung zwischen Objekten einer gewählten Objektklasse und zugehörigen Bildinhalten einer Videoaufnahme. Um dieses äußerst schwere Problem anzugehen wird für gewöhnlich das sogenannte Tracking-durch-Detektionen Paradigma verwendet: Als Erstes werden für die gesamte Aufnahme Objektdetektionen erzeugt. Im zweiten Schritt werden die Zuordnungen zwischen Detektionen und Objekten berechnet. Das Finden der korrekten Zuordnungen wird als Datenassoziationsproblem bezeichnet. Aus der Lösung ergeben sich die Trajektorien der Objekte.

Da das Tracking-by-Detection-Paradigma sequentiell berechnet wird, können sowohl Fehler des Detektors, als auch falsch bewertete zeitliche Konsistenzen zwischen Detektionen zu Fehlerfortpflanzungen führen. Entsprechend bestimmt das verwendete Datenassoziationsmodell die Genauigkeit der Trajektorien wesentlich. Um möglichst genaue Resultate zu erreichen, fokussiert sich diese Arbeit auf die Erforschung robuster Datenassoziationsmodelle. Gleichzeitig werden Standard-Detektoren verwendet, um die entsprechenden Ergebnisse aussagekräftig mit vorherigen Ansätzen vergleichen zu können.

Viele der etablierten Verfahren nutzen für die Datenassoziation lediglich die zeitliche Konsistenz zwischen Detektionen aus, welche in einer Trajektorie direkt aufeinanderfolgen. Solch einfache Modelle sind jedoch stark anfällig gegenüber den erwähnten Fehlern. In dieser Arbeit werden robustere Tracking-Verfahren durch die Verwendung von Datenassoziationsmodellen höherer Ordnung (Higher-Order Multiple Object Tracking) präsentiert. Dabei tragen nicht nur aufeinanderfolgende, sondern alle Paarungen von Detektionen, welche einer Trajektorie zugeordnet werden, zur Bewertung der Konsistenz einer Trajektorie bei. Um die Gesamtinformationen einer Videoaufnahme umfassend auszunutzen, formuliert diese Arbeit die Datenassoziationsmodelle als globale Optimierungsprobleme. Entsprechend werden die Aufnahmen offline, unter Verwendung aller verfügbaren Informationen verarbeitet. Die zugehörigen Optimierungsprobleme sind jedoch $\mathcal{NP}$-schwer. Entsprechend ist es schwierig, gute Lösungen zu berechnen. Zu jedem vorgeschlagenen Datenassoziationsmodel wird ein passendes Lösungsverfahren präsentiert, welches in der Praxis je nach verwendetem Model nahezu optimale oder sogar (nach unserem Wissen erstmalig) beweisbar global optimale Ergebnisse liefert.

Im ersten Teil dieser Arbeit wird ein Verfahren zur verbesserten Ausnutzung der zu einem Zeitpunkt bereitstehenden Signale präsentiert. Während der Standard-Tracking-Ansatz nur Objektdetektionen verwendet, erlaubt die vorgeschlagene Methode, mehrere Eingangssignale global für die Tracking-Aufgabe zu fusionieren. Ein Datenassoziationsmodel höherer Ordnung wird vorgeschlagen, bei dem die Konsistenz sowohl innerhalb eines Signals, als auch zwischen verschiedenen Signalen bewertet wird. Durch die Verwendung komplementärer Signale können Schwächen einzelner Signale kompensiert und Vorteile kombiniert werden. Das vorgeschlagene Datenassoziationsmodell basiert auf einem $\mathcal{NP}$-schweren gewichteten Graph-Labeling Problem. Auf Grund der Komplexität des Problems ist die Berechnung einer Optimallösung schwierig. Es wird ein dafür passendes, approximatives Optimierungsverfahren für das Graph-Labeling Problem vorgestellt. Die Evaluierungen zeigen, dass damit nahezu optimale Lösungen

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

erzeugt werden. Der Nutzen der Fusionierung wird anhand von zwei Anwendungen der Graph-Labeling Formulierung analysiert. (i) Um mehr Bildinformationen auszunutzen, werden Personendetektionen mit Kopfdetektionen kombiniert. Es zeigt sich, dass durch die Fusion wesentlich bessere Ergebnisse als bei ausschließlicher Verwendung von Personendetektionen erreicht werden. Insbesondere hilft es falsch-positive Personendetektionen zu erkennen und entfernen, da diese oft keine passenden Kopfdetektionen haben. Ferner führt der Fusionierungssansatz zu einer längeren Verfolgung von Personen, da im Falle von fehlenden Personendetektionen die Kopfdetektionen zur Lokalisierung und Verfolgung der Personen verwendet werden. (ii) Es werden Inertialmesseinheiten (IMUs) mit einer Videoaufnahme fusioniert. Dazu wird angenommen, dass jede zu verfolgende Person eine IMU am Rücken trägt. Es werden die Beschleunigungs- und Orientierungsmessungen der IMUs mit entsprechend aus der Videoaufnahme geschätzten Werten gekoppelt. Das entsprechende Graph-Labeling Problem erzeugt Trajektorien, welche sowohl zeitlich konsistent bezüglich der Videoaufnahme, als auch zu den zugehörigen IMU-Signalen sind. Die Fusionierung führt insbesondere dann zu erheblich besseren Trajektorienergebnissen gegenüber rein videobasierten MOT-Verfahren, wenn die visuellen Informationen beeinträchtigt sind (etwa durch Bewegungsunschärfe sowie bei ähnlich gekleideten Personen). Fehlende Detektionen lassen sich durch die Fusionierung sehr robust rekonstruieren, sodass die Methode eine geringere Abhängigkeit von der Qualität der Detektionen, verglichen mit rein videobasierten Verfahren, aufweist. Darüber hinaus ermöglicht die Fusionierung, Personen im Bild zu identifizieren, da jede Trajektorie einer IMU zugeordnet wird. Insgesamt zeigen die Methoden aus dem ersten Teil dieser Arbeit, dass die vorgeschlagene Fusionsformulierung es ermöglicht, bereitgestellte Daten umfassender zu nutzen, da mehr der vorhandenen Bildinformationen und Signale integriert werden können. Dadurch verbessert sich die Tracking-Genauigkeit erheblich.

Nichtsdestotrotz stellen Objektdetektionen wertvolle Informationen bereit, welche durch bestehende Verfahren nicht komplett ausgenutzt werden. Datenassoziationsmodelle höherer Ordnung werden auf Grund der Komplexität entweder gar nicht verwendet oder basieren auf heuristischen Optimierungsverfahren, wodurch falsche Zuordnungen erzeugt werden können.

Im Gegensatz dazu wird im zweiten Teil dieser Arbeit ein Optimierungsverfahren vorgestellt, welches es erstmalig ermöglicht, ein dafür passendes Datenassoziationsmodell höherer Ordnung mittels globaler Optimierung zu lösen, wodurch sich insbesondere Langzeitinformationen und -interaktionen ausnutzen lassen, obwohl das Problem $\mathcal{NP}$-schwer ist. Dazu wird ein neues Datenassoziationsmodell vorgeschlagen und als ein binäres lineares Programm beschrieben. Es werden effiziente Separierungsalgorithmen vorgestellt, um das Optimierungsproblem mittels eines Schnittebenenverfahrens zu lösen. Durch die globale Optimierung konnte das Verfahren auf allen getesteten Datensätzen den Stand der Technik erheblich verbessern. Außerdem zeigen durchgeführte Experimente, dass die Methode wesentlich von der Verwendung von Langzeitinformationen profitiert. Die vorgestellte Methode führt auf den verwendeten Datensätzen zu einer nahezu optimalen Zuordnungsgenauigkeit bei gegebenen Detektionen. Zukünftige Arbeiten können sich daher auf andere Bereiche, wie der genaueren Extraktion von Detektionen konzentrieren. Insgesamt zeigt der zweite Teil dieser Arbeit, dass die verbesserte Ausnutzung der über die Zeit bereitgestellten Informationen zu einer erheblichen Verbesserung der Trajektorienergebnisse führt. Erstmalig wurde ein globales

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Optimierungsverfahren zur Lösung von Datenassoziationsmodellen höherer Ordnung erfolgreich eingesetzt.

Schlagwörter: Verfolgung mehrerer Personen, Video, Zuordnungsmodelle höherer Ordnung, Sensorfusion, Binäres lineares Programm, Binäres quadratisches Programm, Global optimale Lösung

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

1 Introduction

Computer systems capable of sensing and interpreting their environment have the potential to positively impact human lives in numerous areas, for instance road safety, as a tool to support urban planning, or in human-computer interaction.

A prerequisite towards obtaining a high-level understanding of the environment is the localization of all objects of a specified object class and the retrieval of their movements using a numerical representation. This task is commonly known as *Multiple Object Tracking* (MOT) or *Multiple People Tracking* (MPT) if only humans are to be tracked.

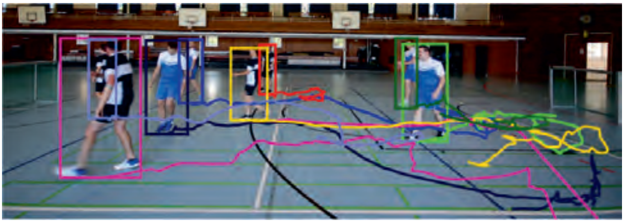


Figure 1.1: Exemplary application of multiple object tracking: Tracking soccer players for performance analyses (see also Section 3.5).

1.1 Applications

A computer system that tracks multiple objects simultaneously has a wide range of applications, as briefly highlighted in this section.

Autonomous Driving. Enabling cars to drive autonomously is considered to be a major improvement to our society [1]. A necessity to achieve this goal is to understand the surrounding of a car. This includes being able to track other cars and people. Therefore, tracking objects filmed by a car is a relevant and ongoing research topic [2]. High tracking accuracies are required. At the same time, a fast computation is needed in order to react quickly. The information obtained from trajectories is used to predict future movements of other road users and, based on that, decide next actions [3]. Thus, the number of accidents is expected to be reduced drastically. Several approaches perform

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

tracking based on *Radio Detection And Ranging* (RADAR) or *Light Detection and Ranging* (LIDAR) signals [4]. There exist also methods based on video information [5, 6]. However, tracking cars in a video signal is challenging as different vehicles may look very similar [7].

Video surveillance. Using MOT, various security-related applications can be implemented. In a surveillance setup, MOT enables the detection of abnormal behavior [8–10]. When tracking cars, accidents or stalled vehicles can be detected as well as cars moving in unusual directions or with excessive speeds. In an epidemic, when social distancing is an important measure to limit the spread, an MPT method can be crucial to detect misconduct [11], take appropriate actions and draw conclusions accordingly.

Object counting. MOT allows counting the number of objects entering, leaving, or being present in a specified area. When applied to a street scene, MOT provides quantitative data about road utilization. This can be used to detect traffic jams or lead to necessary measures such as road extensions or putting up traffic lights. MOT may also deliver detailed information on the number and the behavior of shop customers [12], providing useful and necessary information for example when optimizing the placement of products. Also, tracking information can be used for urban planning [13], *e.g.*, to construct new train stations [14] that better satisfy the needs and behaviors of citizens.

Vehicle speed estimation. Using a static and calibrated camera, the velocity of a car, described in *three-dimensional* (3D) world coordinates, can be estimated by tracking methods [10]. These cost-effective alternatives to RADAR-based approaches enable speed limit enforcement.

Sports. For different areas in sports, tracking athletes (see Figure 1.1) provides useful insights that are either presented to the viewers during a sports broadcast or used to analyze and optimize the performance of athletes [15–19].

Social behavior. Several works have studied human behavior, such as group dynamics [20] and social force models [21]. It has been shown that integrating such models into the tracking formulation yields more accurate trajectory results [22, 23]. In turn, MPT can be used to collect data that allows studying and forming behavior models [14, 24–28].

Animal tracking. Observing the behavior of animals is of great interest to biologists. While animals with a high activity radius are typically tracked independently by a *Global Positioning System* (GPS) device attached to the animals, *e.g.*, birds [29], such an invasive approach is undesirable, and it is often not applicable to smaller animals. Thus, several tracking methods have been developed to compute trajectories for animals based on a video signal [30]. For instance, by using a camera, it is possible to track bees and, based on that, study the waggle dance [31–33]. Also, analyzing fish behavior has a long tradition so that different trackers exist, using *two-dimensional* (2D) [30, 34]

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

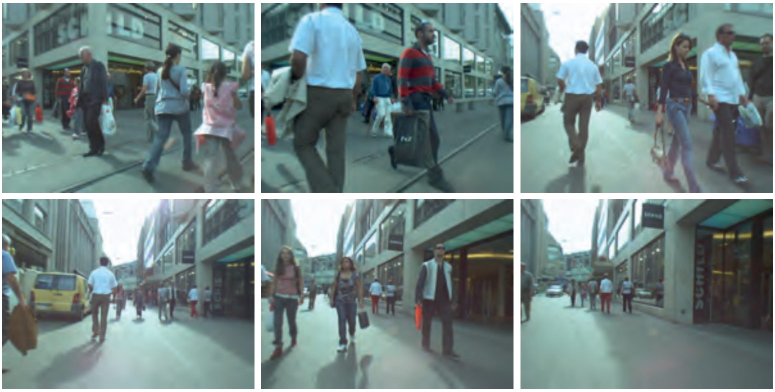


Figure 1.2: Example frames from a test sequence [43, 44] that was used to evaluate the tracking methods of this work. The images are ordered chronologically from left to right and from top to bottom.

or 3D [35] data. Other works focus on tracking ants [36] or bats [37]. Difficulties in video-based tracking of animals arise from high similarities of appearance and frequent occlusions.

Microbiology. Multiple object tracking has also drawn attention in microbiology. For example, it is possible to track swimming microorganisms from microscopic data [38]. Based on that, motion patterns of organisms can be detected [39]. Closely related to tracking is *lineage tracing* [40–42], the tracking of living cells. Here, the setting differs from the standard tracking setup, as cells may divide. The task is to reconstruct so-called lineage trees that encode which cells originate from another cell.

1.2 The Multiple Object Tracking Problem

The task of *Multiple Object Tracking* (MOT) is to localize all objects of a specified class, *e.g.*, persons or cars, and to capture their movements. Both subtasks use measurements from sensor data. By assigning each localization to its corresponding object identity, the *trajectory* of each object is formed. Thus, a trajectory describes the location of a corresponding object as a function of time.

Various sensor types can be used to perform the task such as LIDAR [4] or RADAR [45]. Application scenarios based only on these sensors are limited, though. For instance, LIDAR sensors are relatively expensive[1]. Using LIDAR or RADAR, object occlusions cannot be handled well since re-identification is difficult. In contrast, video camera sensors are cheap, widely used, and allow extracting rich image features that facilitate re-identification of objects after being occluded. Therefore, this work focuses mainly on

---

[1] Prices range from $75.000 down to a few hundred dollars depending on the quality demands and requirements [46].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

video-based tracking. Subsequently, this work considers MOT as the task of computing the locations and trajectories of all objects that appear in a video recording and belong to a determined object class. Figure 1.2 shows parts of a prototypical sequence to which the MOT methods of this work were applied.

In the following, the MOT problem is formalized. First, by a general and abstract definition. Then, the standard procedure to perform MOT is introduced which simplifies the computational costs so that computing trajectories becomes feasible. Finally, the methods are further classified into different model types.

1.2.1 Video-based MOT

An MOT method must return the locations of tracked objects in terms of *bounding-box detections*[2]. A detection describes the location and shape of an object in terms of an axis-aligned rectangle, see Figure 1.3(a). In addition, the MOT method must assign an identity to each of these detections. Ideally, all detections of a sequence that mark the same object obtain the same identity. Accordingly, the detections of an identity result in the trajectory of an object. Figure 1.3(a) demonstrates various challenges an MOT method needs to cope with. Objects may be partially (or fully) occluded. This often causes missing or misaligned detections. Further errors that can be seen in Figure 1.3(a) are double and false positive detections. An MOT method needs to be robust against these types of errors when assigning detections to identities which poses a challenging problem. Also, objects may look similar so that creating a method that decides reliably which detections belong to the same object is challenging. More details about various challenges in MOT are highlighted in Section 1.3. Finally, an example input image with corresponding outputs produced by an MOT method, *i.e.*, object detections and trajectories, is depicted in Figure 1.3(b).

The ultimate goal of MOT is an optimal tracking method. As there is no established standard terminology for such a method in the literature, a formalization is provided subsequently.

Definition 1.1. Let $\mathrm{R} \coloneqq \{1, \dots, n_{\mathrm{R}}\}$ and D be the set of frame indices and the set of detections of a recording, respectively.

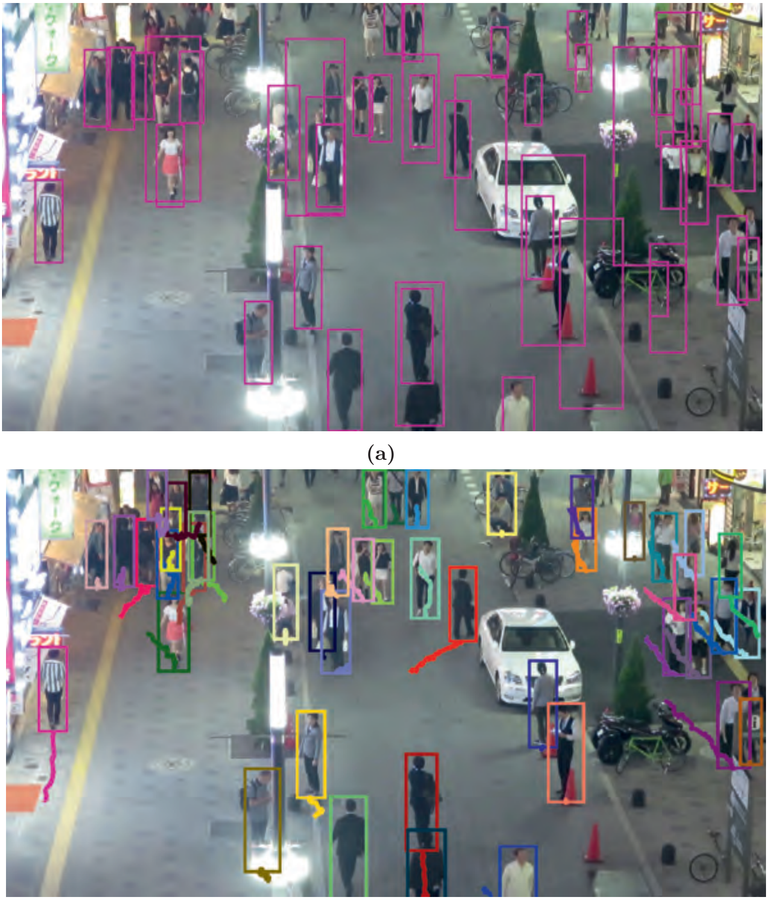
(a) A map $\mathsf{A} : D \to \{0\} \cup \{1, \dots, n_{\mathrm{obj}}\}$ of detections to $n_{\mathrm{obj}} \in \mathbb{N}$ identities is called *quasi data association* if $\mathsf{A}^{-1}(\{n\}) \neq \emptyset \ \forall n \in \{1, \dots, n_{\mathrm{obj}}\}$.

(b) Let $D_{\mathfrak{f}} \subseteq D$ denote all detections in frame $\mathfrak{f} \in \mathrm{R}$ and let $o \in \{1, \dots, n_{\mathrm{obj}}\}$ be an object identity. The set $\gamma \coloneqq \gamma_o \coloneqq \mathsf{A}^{-1}(\{o\})$ is a *trajectory* for object o if $|\gamma \cap D_{\mathfrak{f}}| \leq 1 \ \forall \mathfrak{f} \in \mathrm{R}$, *i.e.*, γ contains at least one detection of object o, and at most one detection of object o per frame. Accordingly, the trajectory of γ at frame $\mathfrak{f}$ is given by $\gamma(\mathfrak{f}) \coloneqq \gamma \cap D_{\mathfrak{f}}$.

(c) A quasi data association is called *data association* if the preimage $\gamma_o = \mathsf{A}^{-1}(\{o\})$ is a trajectory for all objects $o \in \{1, \dots, n_{\mathrm{obj}}\}$.

(d) The number of frames n_{R} is the *length* of a recording.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a)

(b)

Figure 1.3: (a) Video frame of the MOT17 test set [44] with superimposed input detections from the DPM detector [47]. (b) Resulting trajectories using the tracking method Lif_T of Chapter 4.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Remark 1.2. Any detection mapped to 0 by A is considered a false positive detection and is suppressed for the tracking result.

The following definition then categorizes an MOT method.

Definition 1.3. For a recording, a *multiple object tracking* method returns

(a) a finite set D of bounding-box detections,

(b) a data association $\mathsf{A} : D \to \{0\} \cup \{1, \ldots, n_{\text{obj}}\}$ to n_{obj} identities, where $n_{\text{obj}} \in \mathbb{N}$.

The objective is to create the best possible MOT result, which may be measured in terms of the errors produced by an MOT method. Since errors may occur both at the localizations (*e.g.*, by failing to detect an object) and the assignments to identities (*e.g.*, by assigning a detection to a wrong identity), assessing the quality of an MOT system is not straightforward; various metrics have been proposed that all reflect different characteristics of a tracking result. They are introduced and discussed in Section 2.7.4.

The *optimal* MOT result is a set of detections D that describe all appearing objects and the data association A groups all detections to their respective identities. To formalize an optimal MOT method, let n_{obj} denote the number of all individual objects of a specified object class that appear in a recording. For object $o \in \{1, \ldots, n_{\text{obj}}\}$ and frame $\mathfrak{f}$, we assume a labeling function $l(o, \mathfrak{f}) \in \{0, 1\}$ to output 1 if object o is within the captured image space at the respective frame, and 0 otherwise.

Definition 1.4. Let $\mathrm{R} = \{1, \ldots, n_{\mathrm{R}}\}$ be the set of frame indices of a recording. A multiple object tracking method is *optimal* if:

(a) There is a bijection b from $\{(o, \mathfrak{f}) \in \{1, \ldots, n_{\text{obj}}\} \times \mathrm{R} \mid l(o, \mathfrak{f}) = 1\}$ to D such that $b\left((o, \mathfrak{f})\right) =: \mathbf{d} \in D$ if and only if detection $\mathbf{d}$ localizes[3] object o in frame $\mathfrak{f}$.

(b) The data association A of the MOT method satisfies for all $\mathbf{d}_1, \mathbf{d}_2 \in D$, $\mathsf{A}(\mathbf{d}_1) = \mathsf{A}(\mathbf{d}_2)$ if and only if $\mathbf{d}_1$ and $\mathbf{d}_2$ correspond[3] to the same identity.

Hence, an optimal MOT method localizes each object by exactly one detection box according to Definition 1.4(a). Moreover, if Definition 1.4(b) holds, the data association assigns each object correctly to its identity.

Optimally solving the MOT problem is demanding as it includes the two tasks object detection (which is a research field on its own [48, 49]) and data association (which often leads to difficult combinatorial problems [50–55]). These two problems pose correlated tasks: detections apparently influence the data association. Conversely, the data association may reflect information about the presence of objects in time and space. Some works exploit this relation by formulating an optimization problem that simultaneously optimizes both tasks [56–60].

---

[2] A formal definition of a bounding box detection is provided in Section 2.7.

[3] There are different definitions for this term, each resulting in a different assessment of a tracking method, see Section 2.7.4.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

1.2.2 Tracking-by-detection

The predominant approach to tackle the MOT problem is to consider the two necessary outputs of Definition 1.3 as consecutive tasks [54, 61–67]. Such a procedure is called the *tracking-by-detection* paradigm [68]. In the first step, an object detector returns for each frame independently the putative locations of all objects, yielding a set of initial detections D' for a sequence. In the second step, these hypotheses are linked across frames to form trajectories, ensuring consistency between all detections of an identity and removing implausible detections. The final result is the set of remaining detections $D \subset D'$ and corresponding data association A.

By decoupling the two tasks, greater flexibility is achieved. Any object detector that outputs detections for the considered object class may be used for the tracking-by-detection paradigm. Technical progress in object detection therefore usually leads to improvements of an MOT method. At the same time, advances on the data association part are developed that are often independent of a specific object detector. Accordingly, the two tasks can be considered as (nearly) independent research fields.

Within the scope of tracking-by-detection, the goal is to compute the best trajectories conditioned on pre-computed detections. Here, *best* is understood probabilistically as a maximum a-posteriori problem [68].

Definition 1.5. The MOT task using the tracking-by-detection paradigm is to find the trajectories that maximize the posterior probability w.r.t. a probability measure P, given all observed input detections D, formally:

$$\widehat{\Gamma} = \arg\max_{\Gamma} \mathsf{P}(\Gamma \mid D)\,, \tag{1.1}$$

where the optimization is performed over all trajectories Γ that can be created from the detections D.

Thus, for a given set of detections D, the quality of an MOT outcome depends entirely on the chosen probability model P and how well the problem defined by Eq. (1.1) can be maximized.

Solving Eq. (1.1) is infeasible in practice since the search space of problem (1.1) is huge and an efficient solver for the corresponding optimization problem does not exist. Therefore, in the following, assumptions on the search space and probability model are made to simplify the computations.

Assuming independence in the probability model. Assuming pairwise independence of the trajectories Γ and conditional independence of the detections D given the trajectories Γ, the solution $\widehat{\Gamma}$ to the maximum a-posteriori problem of Eq. (1.1) for a

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

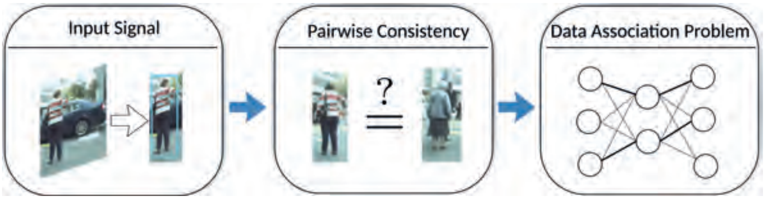


Figure 1.4: Visualization of the tracking-by-detection paradigm. First, detections are generated. Then, pairwise costs between detections are computed. Finally, the information is used in a graph model that represents all feasible connections between detections. Typically a discrete optimization problem is solved to obtain the trajectories. The data association needs to compensate for errors introduced by the preceding steps. Some parts of the image are taken from the MOT16 dataset [44].

sequence of length n_R can be obtained[4] following Zhang *et al.* [61] as:

$$\begin{aligned}
\widehat{\Gamma} &= \arg\max_{\Gamma} \mathsf{P}(D \mid \Gamma)\mathsf{P}(\Gamma)\,, \qquad (1.2)\\
&= \arg\max_{\Gamma} \prod_{\mathfrak{f}=1}^{n_\mathrm{R}} \mathsf{P}(D_\mathfrak{f} \mid \Gamma)\mathsf{P}(\Gamma)\,,\\
&= \arg\max_{\Gamma} \prod_{\mathbf{d}\in D_1\cup\cdots\cup D_{n_\mathrm{R}}} \mathsf{P}(\mathbf{d} \mid \Gamma) \prod_{\gamma\in\Gamma} \mathsf{P}(\gamma)\,,\\
&= \arg\min_{\Gamma} -\sum_{\mathfrak{f}=1}^{n_\mathrm{R}} \sum_{\mathbf{d}\in D_\mathfrak{f}} \log(\mathsf{P}(\mathbf{d} \mid \Gamma)) - \sum_{\gamma\in\Gamma} \log(\mathsf{P}(\gamma))\,.
\end{aligned}$$

The imposed simplifications ignore any dependencies between trajectories. An optimal solution corresponds to the most plausible trajectories that are consistent with the observed detections. The difficulty of problem (1.2) is analyzed in Section 1.3.3, showing that the optimization problem is challenging. Accordingly, the more general problem (1.1) is indeed challenging, too.

Reducing MOT to pairs of detections. To further reduce the complexity of Eq. (1.2), it is a common approach to assume that $\mathsf{P}(\gamma)$ can be computed based on probabilities between pairs of detections belonging to γ, see *e.g.*, Zhang *et al.* [61]. Some models additionally use start and end probabilities of a trajectory. As in Eq. (1.2), probabilities are converted into negative log-likelihoods, which we call *costs*. They are further categorized into *pairwise* costs corresponding to probabilities between pairs of detections and *unary* costs corresponding to probabilities of detections.

An MOT method based on the tracking-by-detection paradigm using pairwise costs performs mainly three steps. First, an object detector is applied to a video recording to obtain hypotheses about the presence of objects of a specified object class. Then, unary and pairwise costs are computed based on the input detections. Finally, an optimization

---

[4] The reformulation as a minimum makes it applicable to the optimization methods presented in Section 2.6.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

problem is solved to obtain the most likely trajectories given the object hypotheses and likelihoods. Figure 1.4 illustrates these steps.

The last step is typically computed within a *data association graph*[5] that models all feasible trajectories given the object hypotheses and associates each trajectory with a corresponding cost value. If a *node* (corresponding to a detection) or an *edge* (representing a possible link of two nodes to a trajectory) is part of the solution, the respective costs are summed up in the objective function.

One way to simplify $\mathsf{P}(\gamma)$ based on pairwise costs is to assume that $\mathsf{P}(\gamma)$ depends only on the pairwise probabilities of consecutively linked detections of a trajectory. Thus, the probabilistic model assumes that trajectories obey the Markov chain assumption [69]: Given a family of random variables $Y = (X_i)_{i \in \mathbb{N}}$, where each random variable X_i has the same finite set S as measurement space, then Y obeys the *first-order Markov chain* assumption if

$$\mathsf{P}(X_n = s_n \mid X_{n-1} = s_{n-1}, \dots, X_1 = s_1) = \mathsf{P}(X_n = s_n \mid X_{n-1} = s_{n-1}) \tag{1.3}$$

for all $s_1, \dots, s_n \in S$ and for all $n \in \mathbb{N}$. In this case, the probability of a trajectory $\gamma = \{\mathbf{d}_1, \dots, \mathbf{d}_M\}$ using the chain rule and the Markov property is given by

$$\mathsf{P}(\gamma) = \mathsf{P}(\mathbf{d}_1) \prod_{i=2}^{M} \mathsf{P}(\mathbf{d}_i \mid \mathbf{d}_{i-1}) \mathsf{P}(\mathbf{d}_M)\,, \tag{1.4}$$

where we assume that detection $\mathbf{d}_i$ is from frame $\mathfrak{f}_i$ and $\mathfrak{f}_i < \mathfrak{f}_j$ if $i < j$. Thus, the probabilistic model assumes that the assignment of a detection $\mathbf{d}_i$ to a trajectory γ is only stochastically dependent on $\mathbf{d}_{i-1}$. An MOT approach based on assumption (1.3) is a *Markovian* multiple object tracking method or a *first-order* MOT method. The corresponding optimization problem can be solved efficiently to global optimality [61]. A Markovian data association graph is depicted in Figure 1.5. However, the tracking accuracy deteriorates due to an oversimplified model.

Higher-order MOT. More expressive MOT models exploit not only the local information between detections that are directly linked within a trajectory but also the global consistency of an entire trajectory, which we call *Higher-Order Multiple Object Tracking* (HO-MOT). Problem (1.2), without further simplifications, is an *explicit* HO-MOT formulation, as the consistency of an entire trajectory is explicitly modeled. So far, explicit HO-MOT is of limited use in practice, as the underlying optimization problem is too challenging to be solved, and in addition, the mapping of each possible trajectory to a probability value is difficult. An alternative approach is to employ a non-Markovian MOT formulation that assesses trajectories by incorporating all available pairwise probabilities within a trajectory. This enables to exploit implicitly long-range temporal interactions. Since such a model reasons only from pairwise costs, we call it *implicit* HO-MOT. Figure 1.6 shows a prototypical situation that is often not solved correctly by first-order MOT methods, in contrast to higher-order approaches. Using an HO-MOT formulation helps to improve the precision of tracking results, see Figure 1.7. For implicit HO-MOT, it is crucial to include many pairwise probabilities into the decision process, especially long-term connections to take long-range temporal interactions

---

[5] Terminologies of graph theory are briefly summarized in Section 2.3.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

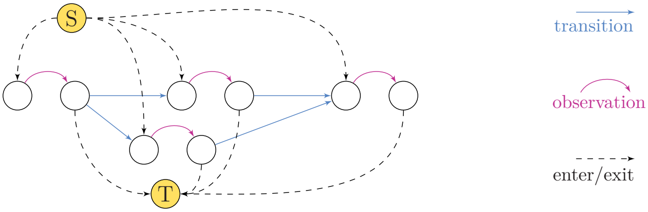


Figure 1.5: Data association using a network flow graph [61]. The start and end of a trajectory are encoded by an *enter* edge from node S and an *exit* edge to node T, respectively. Each detection of a trajectory corresponds to an *observation* edge. The linking of two detections between different frames to a trajectory is represented by a *transition* edge. All edges have costs indicating the likelihood for the respective event. The number of selected incoming edges at a node must equal the number of selected outgoing edges, except for nodes S and T. Transition edges skipping frames are not drawn for visualization purposes.

into account. At the same time, long-term connections allows for the compensation of detector errors in cases where objects are temporarily missed, see Figure 1.8. While an HO-MOT approach (implicit and explicit) is theoretically preferable to a Markovian MOT formulation, the underlying optimization problem is challenging, as it usually leads to an $\mathcal{NP}$-hard optimization problem [52, 54, 63, 64, 70]. Existing methods deliver only heuristic HO-MOT solutions so that the tracking accuracy deteriorates due to non-optimal decisions.

Connectivity priors. With longer temporal distances between detections, corresponding pairwise costs become less discriminative and unreliable. This becomes apparent if a motion model is employed to define pairwise costs. When spatial distance is considered between detections that are temporally far apart, many of the possible connections will be plausible. At the same time, there may be connections that can be clearly classified as false, *e.g.*, a connection that represents a motion that would be too fast for an object to perform due to physical constraints.

In order to leverage information of long temporal distances, an extension to standard data association graphs is the introduction of additional *lifted* edges that induce path connectivity priors [50, 64]. A lifted edge between nodes (representing detections) ensures when activated that a trajectory on the original data association graph (called *base graph*) has to connect the two ends of the respective lifted edge and vice versa. Thus, it augments the underlying tracking model by additional connectivity priors, resulting in a more expressive tracking model. Two settings are of particular interest:

(i) Base edges connect nodes over short temporal distances while lifted edges connect nodes over long temporal distances. This allows taking long-term considerations reliably into account. If a lifted edge indicates that matching the respective two detections is plausible, there must still be a consistent trajectory connecting the end nodes of the lifted edge. Conversely, if a lifted edge indicates that two detections very likely contain

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

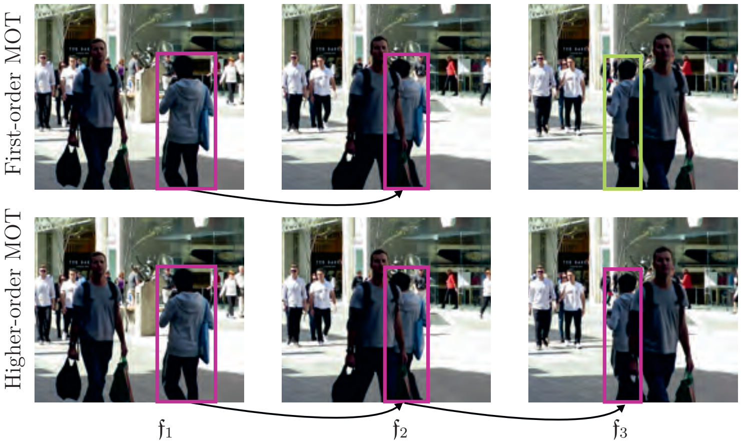


Figure 1.6: Comparison of first-order MOT *vs.* higher-order MOT on a benchmark sequence [44]. The person detected in frame $\mathfrak{f}_1$ is fully visible. Due to partial occlusion, only the right and left side of the person is visible in frame $\mathfrak{f}_2$ and $\mathfrak{f}_3$, respectively. As there is sufficient overlap of the image contents, the detections of frame $\mathfrak{f}_1$ and $\mathfrak{f}_2$ are linked by an appearance-based first-order MOT method. However, there is little overlap between the image contents defined by the detections in frame $\mathfrak{f}_2$ and $\mathfrak{f}_3$. Also, the background has greatly changed. Consequently, a first-order appearance-based MOT method erroneously creates a new trajectory (indicated by the green detection). In contrast, a higher-order MOT method exploits that (i) the detection in frame $\mathfrak{f}_1$ shares appearance similarities with the detection in frame $\mathfrak{f}_2$ and $\mathfrak{f}_3$ since both depict one side of the person and that (ii) the detections in frame $\mathfrak{f}_2$ and $\mathfrak{f}_3$ do not completely rule out a common trajectory. As a result, an HO-MOT method assigns all detections to the same person.

different objects, there must not be a trajectory connecting the corresponding nodes. A prototypical situation demonstrating the two cases is shown in Figure 1.9. A graph with lifted edges is depicted in Figure 1.10. (ii) Each base edge is duplicated by a lifted edge. In this case, connectivity priors induce an implicit HO-MOT model, as shown in Figure 1.11.

The existing MOT formulation [64] with lifted edges poses a challenging $\mathcal{NP}$-hard optimization problem so that existing methods rely on heuristic solvers [51, 64]. Accordingly, the tracking accuracy is deteriorated.

1.3 Challenges of Multiple Object Tracking

Although MOT has been a focus of research for decades [71, 72], it still poses a challenging problem. We discuss the corresponding challenges for all three steps of the tracking-by-detection paradigm (see Figure 1.4).

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

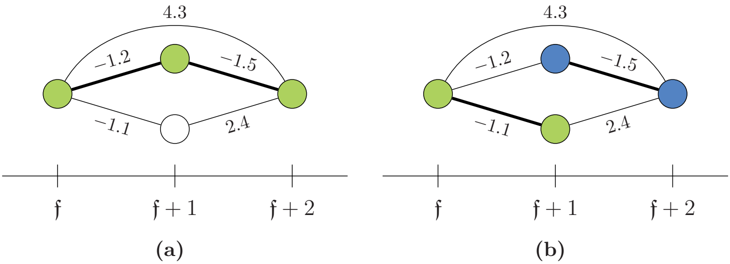


Figure 1.7: Optimal trajectories according to different data association models. Both figures depict the same data association graph with all feasible detection assignments between the frames $\mathfrak{f}$, $\mathfrak{f}+1$, and $\mathfrak{f}+2$, together with corresponding pairwise costs. Nodes of the same color form respective trajectories. (a) The Markovian model considers only costs between directly linked detections of a trajectory. Thus, the optimal trajectory has an objective value of $(-1.2)+(-1.5) = -2.7$. It ignores the edge from frame $\mathfrak{f}$ to frame $\mathfrak{f}+2$ that signals that the first and last detection belong to a different object. (b) An implicit HO-MOT model takes all pairwise probabilities within a trajectory into account. Consequently, it prevents the creation of the wrong tracking result.

1.3.1 Errors caused by the object detector

The first step within a tracking-by-detection approach, the object detection, has seen tremendous improvements with the rise of neural networks in computer vision [73]. This becomes apparent when comparing a traditional method, depicted in Figure 1.3(a), with detections from a neural network [73] in Figure 1.12. However, detectors are still far from being perfect.

False negative detections. Objects are frequently missed in the case of (partial) occlusion, as shown in Figure 1.12. Missing detections can also occur when objects are far away from the camera, as objects then appear small in the projected image. Also, persons in rare poses might be missed. Such severe errors are not directly solvable by employing a better data association graph or using a more complex probability model, as there is no input to be assigned. Instead, if a detection at a time before an occlusion and after an occlusion are assigned to the same object, the missing detection can be potentially recovered, *e.g.*, using linear interpolation. Accordingly, it is crucial that the employed MOT model incorporates not only edges connecting detections between consecutive frames but also edges that skip frames.

Misaligned detections. Multiple objects to be tracked might be contained within a single detection box. This frequently happens if objects are close together or if one object is partially occluding another object, as shown in Figure 1.13(a). As a consequence, appearance information might be misleading. Since each detection box is assumed to represent at most one object, the error case is not correctly representable

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

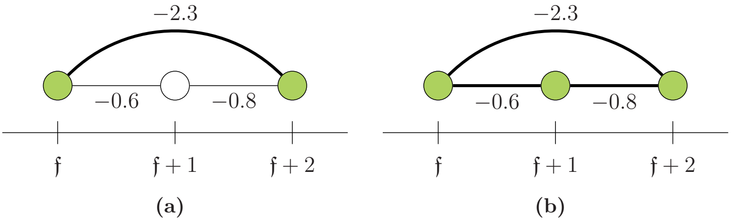


Figure 1.8: Improvement in recall when using an implicit HO-MOT model over a Markovian MOT formulation. Incorporating connections that skip some frames is a prerequisite for correctly creating trajectories for objects that are missed by the detector for some frames. (a) In the Markovian setting, at most one detection from a later image frame is selected for each detection. Consequently, all possible connections are competing with each other so that long-term temporal edges might deteriorate the tracking quality in the Markovian setting by mistakenly skipping detections. In the depicted example, the thick edge forms a trajectory resulting from a Markovian MOT model that skips frame $\mathfrak{f}+1$. (b) In contrast, an implicit HO-MOT model groups all three detections correctly together, as all pairwise costs are negative (indicating that all detections likely belong to the same object).

and resolvable by the tracking-by-detection paradigm.

False positive detections. Detectors might create detection boxes that do not localize the desired object, *i.e.*, false positives, see Figure 1.3. An MOT method thus needs to recognize and remove these detections, which is particularly difficult if the detector consistently creates false positives, *e.g.*, at a background object.

1.3.2 Challenges in discriminative features

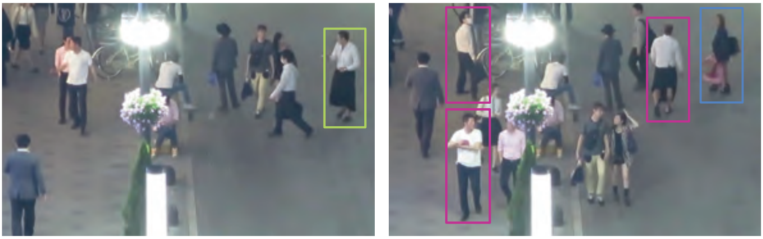
By assuming that the probability of a trajectory is decomposable into pairwise probabilities, the probability measure P needs to reflect the probability of two detections belonging to the same person.

1.3.2.1 Pairwise features.

Given a pair $e = \{v, u\}$ of two detections that might belong to the same object, a vector $\boldsymbol{\psi}(e) \in \mathbb{R}^n$ is constructed, called *pairwise feature* or *pairwise affinity*. A discriminative (pairwise) feature distinguishes pairs of detections belonging to the same object from detections of different objects in terms of its vector representation. A feature vector can be transformed into probability $p_{\boldsymbol{\psi}(e)} \in [0, 1]$ using logistic regression [74].

In order to decide whether two detections belong to the same identity, most tracking methods employ affinities that can be categorized as either processing geometric information based on the corner coordinates of a detection box or visual information within

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a) (b)

Figure 1.9: Lifted edges enable to incorporate long-term temporal information. Two images of an MOT benchmark [44] with a temporal distance of 5.5 seconds are shown. (a) The green rectangle marks a person that shall be tracked. (b) Purple detections indicate high appearance similarities to the query person. Due to the long temporal distance, pairwise features between the two frames are unreliable. However, lifted edges allow exploiting these correspondences reliably. A lifted edge enforces that each potential match must be supported by a plausible trajectory connecting the detections of the two frames. The blue detection box indicates a person that can be clearly distinguished from the query person. A lifted edge then encodes that each trajectory must not connect the two detections.

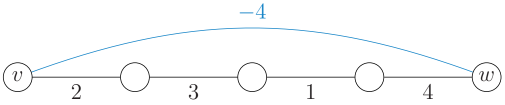


Figure 1.10: Data association graph with base edges and weights (black), and lifted edge and lifted weight (blue). The cost value of the lifted edge indicates that the nodes v and w correspond to the same object. However, this is not consistent with the (short-term) base edges. As a result, the detections are not grouped to a trajectory since summing up all weights gives a worse objective value than not creating a trajectory.

a detection box. However, constructing discriminative features poses a challenging problem on its own, as the features need to cope with many different situations that may appear during a recording.

Geometry-based features, or *spatio-temporal* features, incorporate position and motion cues. For people tracking, these features can be designed specifically to the characteristics of humans, *e.g.*, using a constant velocity assumption [22, 23] or by incorporating social force models [22, 23]. However, the performance of these approaches degrades if motions become more dynamic or people get temporarily occluded. Also, without prior knowledge about the scene or camera calibration, the 3D position information of objects is available only in the form of projected pixel coordinates in the image space which causes ambiguities regarding their positions and motions. If the camera is non-static, measured motions represent movements of objects superimposed by the camera motion. Consequently, spatio-temporal features are robust only within a limited

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

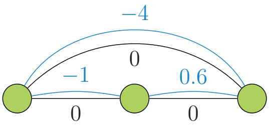


Figure 1.11: If all depicted nodes belong to the same trajectory, then all costs within the trajectory contribute to the objective value of the underlying optimization problem by definition of lifted edges, thus inducing higher-order consistencies. The information given indicates that all nodes belong to the same object. Accordingly, the optimal result is to group all nodes to one trajectory, which is indicated by the green nodes. Base edges (and weights) are drawn in black and lifted edges (and weights) in blue.

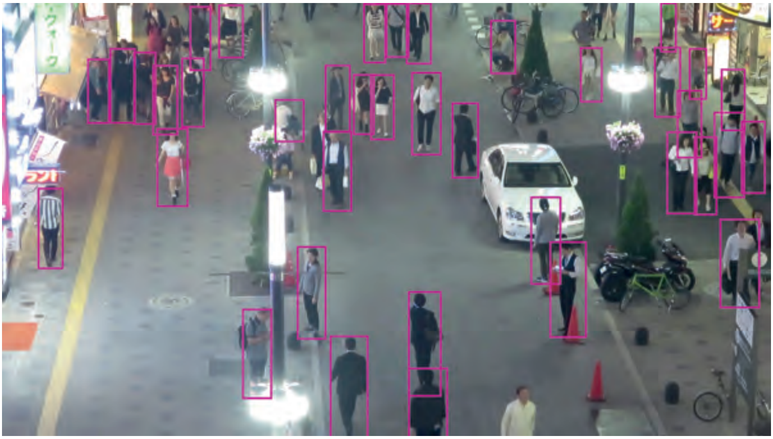


Figure 1.12: The input frame of Figure 1.3(a) with detections from FRCNN [73]. The detection method is introduced in Section 2.7.

temporal range [63].

As a consequence, current state-of-the-art methods employ pairwise features that reason from appearance information [55, 63, 64, 76–82]. In contrast to spatio-temporal cues, appearance features can be applied more reliably to comparatively long temporal distances as they are more robust to camera motions. Also, they do not rely on the constant velocity assumption. Some works thus exploit appearance cues using an object-dependent visual object tracker [81, 83, 84] that requires an initial mask of the object to be tracked and essentially detects the object in each frame. During this process, the appearance model of the object is updated frequently. Also, attention weights [80, 82] have been used to estimate whether the appearance information of a detection box is reliable. However, the accuracy of appearance-based trackers still deteriorates under different frequently occurring effects, *e.g.*, lighting or camera perspective changes, partial occlusions, tiny detection boxes, or if noise is present, see Figure 1.14. Particularly difficult when using appearance features are cases where the apparel of different objects

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

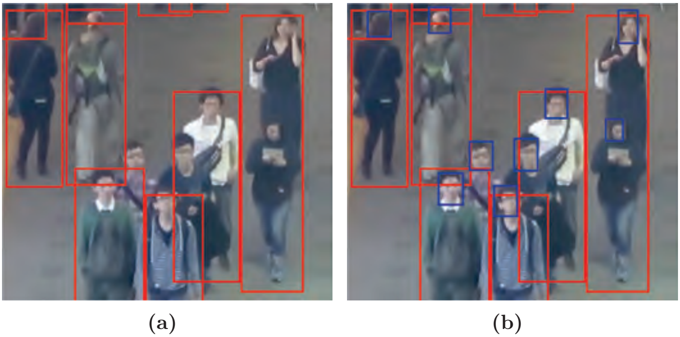


Figure 1.13: Exemplary scene showing detection errors due to partial occlusion. Depicted is a clip of the MOT17-03 sequence [44] with provided input detections (FRCNN). (a) Multiple people appear within a detection box and some persons are not detected at all. (b) The errors can be resolved using, in addition, fine-graded head detections (blue).

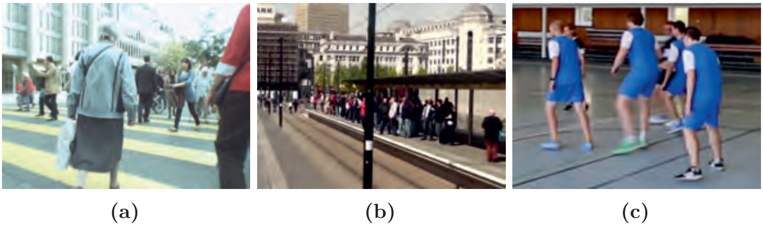


Figure 1.14: Exemplary scenes that show difficult situations for pairwise MOT features. (a) & (b) show two scenes of the MOT16 dataset [44] that are filmed by a moving camera. People are frequently occluded. (c) shows a clip of an image of the VIMPT2019 dataset [75]. Due to the jerseys, distinguishing identities is difficult for a computer vision system.

are very similar, for instance in team sports as depicted in Figure 1.14(c) or when people are changing clothes during a recording. Building accurate pairwise features is thus challenging as they need to cope with all these issues.

1.3.3 Combinatorial challenges

The optimal assignment of detections to identities, as described by Eq. (1.1), poses a discrete optimization problem. Finding a global optimal solution without further simplifications is not feasible in practice from a combinatorial point of view, as no efficient solver for the optimization problem is known. The following shows that even a simplified model (explicit HO-MOT) is difficult to solve, as it belongs to the complexity class $\mathcal{NP}$-hard. Such problems are considered challenging. Even the further simplified

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

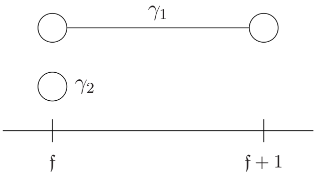


Figure 1.15: Depicted are two trajectories γ_1 and γ_2 with a vector representation $I(\gamma_1) = (1,1)$ and $I(\gamma_2) = (2,0)$, respectively, using an order on the detections accordingly.

implicit HO-MOT models [52, 54, 63, 64, 70] are commonly $\mathcal{NP}$-hard. In particular, no algorithm that solves all instances of an $\mathcal{NP}$-hard problem with polynomial runtime complexity is known[6]. The existence of one such algorithm would imply that many difficult problems could be solved efficiently. We conclude that the MOT problem is challenging.

Complexity of problem (1.2)**.** Assuming all detections to be equally probable, HO-MOT problem (1.2) simplifies to

$$\arg\max_{\Gamma} \mathsf{P}(\Gamma \mid D) = \arg\min_{\Gamma} -\sum_{\gamma\in\Gamma} \log(\mathsf{P}(\gamma))\,. \tag{1.5}$$

Now Eq. (1.5) can be rewritten as a binary linear optimization problem: For a sequence consisting of n_{R} frames, let $n_{\mathfrak{f}} := |D_{\mathfrak{f}}|$ denote the number of detections in frame $\mathfrak{f}$. We consider the bijection I, mapping a trajectory γ to a vector $I(\gamma) = (i_1, \ldots, i_{n_{\mathrm{R}}})$, where $I(\gamma)_{\mathfrak{f}} = i_{\mathfrak{f}}$ denotes the index of the selected detection by γ in frame $\mathfrak{f}$, and equals 0 if $\gamma(\mathfrak{f}) = \emptyset$. An example is given in Figure 1.15. Each non-empty trajectory γ is assigned the cost value $c_{I(\gamma)} := -\log(\mathsf{P}(\gamma))$, while we set $c_{(0,\ldots,0)} := 0$ for an empty trajectory.

Let $N := \{0,\ldots,n_1\} \times \ldots \times \{0,\ldots,n_{\mathrm{R}}\}$ be the set of all possible tuples that encode a trajectory. Then, problem (1.5) is equivalent to the binary optimization problem

$$\min_{z\in\{0,1\}^N} \sum_{i_1=0}^{n_1} \cdots \sum_{i_{n_{\mathrm{R}}}=0}^{n_{\mathrm{R}}} c_{(i_1,i_2,\cdots,i_{n_{\mathrm{R}}})} z_{(i_1,i_2,\cdots,i_{n_{\mathrm{R}}})} \tag{1.6}$$

such that

$$\sum_{i_1=0}^{n_1} \cdots \sum_{i_{\mathfrak{f}-1}=0}^{n_{\mathfrak{f}-1}} \sum_{i_{\mathfrak{f}+1}=0}^{n_{\mathfrak{f}+1}} \cdots \sum_{i_{n_{\mathrm{R}}}=0}^{n_{\mathrm{R}}} z_{(i_1 i_2 \cdots i_{n_{\mathrm{R}}})} = 1\,, \quad \forall \mathfrak{f} \in \mathrm{R}, \forall i_{\mathfrak{f}} \in \{1,\ldots,n_{\mathfrak{f}}\}\,. \tag{1.7}$$

The constraints (1.7) ensure that each detection is assigned to at most one trajectory.

Problem (1.6) is a *multidimensional assignment problem* [85]. In the context of MOT, it is also called *multi-hypothesis* tracker. However, the multidimensional assignment problem is known to be $\mathcal{NP}$-hard for $n_{\mathrm{R}} \geq 3$ [86]. Note also that the solution space has exponential growth. Hence, computing the global optimum is difficult. Besides, it

---

[6] A summary of the field is provided in Section 2.5.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

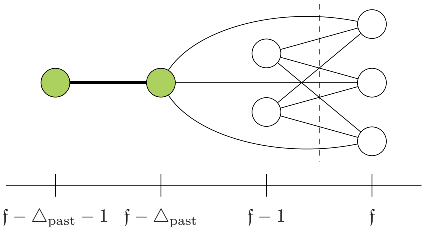

$\mathfrak{f} - \triangle_{\text{past}} - 1$
$\mathfrak{f} - \triangle_{\text{past}}$
$\mathfrak{f} - 1$
$\mathfrak{f}$

Figure 1.16: Data association graph for online tracking at frame $\mathfrak{f}$. Nodes left to the dashed line correspond to unused detections or to already computed trajectories of the past. The nodes from the past may be linked to a detection of frame $\mathfrak{f}$ if they contain at least one detection in frame $\mathfrak{f}' \in \{\mathfrak{f} - \triangle_{\text{past}}, \dots, \mathfrak{f} - 1\}$. The green nodes form an active trajectory that may be connected to a node of frame $\mathfrak{f}$.

is challenging to accurately assign each possible trajectory γ a meaningful cost value c (corresponding to $-\log(\mathsf{P}(\gamma))$. Therefore, solving MOT in terms of Eq. (1.1) or even problem (1.5) is challenging.

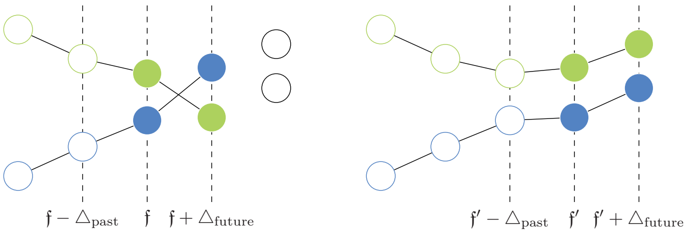
1.4 Related Work

For fixed input detections and pairwise features, the tracking accuracy of an MOT method depends on the chosen data association model and how well the associated optimization problem can be solved. This section presents a literature overview of published data association models in MOT. As there is a vast amount of literature on this topic, only the most relevant works are briefly summarized which significantly impacted subsequent publications in MOT.

Online Tracking. A straightforward way to simplify problem (1.1) is to perform tracking *online* [58, 80–82, 87–95]. In this case, the task is to match the detections of the current frame $\mathfrak{f}$ with trajectories and unused detections of the past. Connections are considered only if the time distance is at most some threshold $\triangle_{\text{past}}$. Trajectories may be matched if their last detection is within the defined time distance. Such a trajectory is *active*, see Figure 1.16.

Some works perform data association heuristically or greedily [80, 81, 89, 90]. The method of Zhu *et al.* [80] propagates each trajectory using an identity-specific tracker until it becomes unreliable (according to a confidence score). A neural network with attention mechanisms is used to identify similar image content between a candidate detection and a detection of a considered trajectory. This information is integrated into a neural network with an attention mechanism and bidirectional long short-term memory (LSTM) components to measure consistency between a detection and an entire trajectory. The assignment of a detection to a trajectory is done greedily, using the best fitting detection (above a similarity threshold).

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.


$\mathfrak{f} - \triangle_{\text{past}}$
$\mathfrak{f}$
$\mathfrak{f} + \triangle_{\text{future}}$
$\mathfrak{f}' - \triangle_{\text{past}}$
$\mathfrak{f}'$
$\mathfrak{f}' + \triangle_{\text{future}}$

(a) **(b)**

Figure 1.17: Data association graph for near-online tracking. (a) The associations within the range $[\mathfrak{f} - \triangle_{\text{past}}, \mathfrak{f} + \triangle_{\text{future}}]$ are computed. Subsequently, all connections between filled circles are reset. (b) Local errors are corrected in the next iteration on frame $\mathfrak{f}' = \mathfrak{f} + 1$.

Different approaches use probabilistic inference, *e.g.*, Kalman Filter [68, 96, 97] or Particle Filter [68, 98, 99] propagate the state of each active trajectory, given the detection information at the current frame and a motion model, which typically includes a constant velocity assumption. Alternatively, several works [58, 82, 91–94] model the data association within a bipartite graph (see Figure 2.4), which can be computed in $\mathcal{O}(n^3)$ in the number of nodes n using the Hungarian method [100].

Such approaches allow for processing the tracking problem efficiently, as the currently captured frame can immediately be used to assign new detections to already existing trajectories or unassigned detections from previous frames. The tracking accuracy heavily depends on discriminative pairwise features, as slightly wrong values in the affinities cannot be compensated by the data association model, potentially leading to error propagation. However, there might be information in some future frames that would enable detecting and preventing such errors, *e.g.*, in the case of an object that is occluded in the current frame. Thus, even if decisions are optimal within the local scope, they may lead to wrong assignments with respect to the tracking task.

Near-Online Tracking. This conceptual deficiency is corrected using so-called *near-online* trackers [65, 101]. For the current frame $\mathfrak{f}$, the approach is to process the information from the past and a limited number of future frames so that the data association is performed within $[\mathfrak{f} - \triangle_{\text{past}}, \mathfrak{f} + \triangle_{\text{future}}]$, resulting in a short delay by $\triangle_{\text{future}}$ frames. While the entire interval is used to find the assignments, only the connections up to frame $\mathfrak{f}$ are updated. Connections to future detections within the interval $[\mathfrak{f} + 1, \mathfrak{f} + \triangle_{\text{future}}]$ are reset after the data association part. The process is then repeated for the subsequent frames, as demonstrated in Figure 1.17. While an accuracy improvement over online models can be expected, valuable information of a recording is still ignored, as not all frames are exploited.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Offline Tracking. The other extreme is to process an entire sequence at once, which is called *offline tracking*, using the information from all frames for the data association. Still, models vary in how they integrate the information into the tracking method. They can be classified into Markovian MOT and HO-MOT.

Markovian MOT. A frequently used MOT approach assumes that the first-order Markov chain assumption holds for the detections of a trajectory [23, 61, 67, 102–106]. Zhang *et al.* [61] show that the optimal MOT solution can then be computed in a flow network with unit capacities by finding the flow of minimal costs. Decisions are guaranteed to be optimal and can be computed efficiently using a min-cost flow algorithm [107, 108] with polynomial runtime.

By assumption, pairwise affinities like distances, speeds, and visual similarities may be incorporated into a Markovian MOT method but no long-term considerations like accelerations. Therefore, it ensures consistencies only between consecutively linked frames.

Higher-order MOT. Some methods directly tackle explicit HO-MOT, *e.g.*, using hyperedges [109–112] that jointly assign probabilities for grouping multiple detections. Even more, multi-hypothesis trackers [53, 79, 113, 114] directly assign to each possible trajectory γ a probability without factorizing $\mathsf{P}(\gamma)$ further (see Eq. (1.6)). This allows to model spatial and temporal consistencies for the whole time span of a trajectory and to incorporate higher-order correlations between the detections. As the underlying multidimensional assignment problem is challenging to solve (see Section 1.3.3), heuristic filtering steps are necessary to reduce the search space. Consequently, it leads to non-optimal solutions that degenerate the tracking accuracy heavily.

In contrast, implicit higher-order models (see Section 1.2.2) that reason from pairwise costs have turned out to be more effective [50, 52, 54, 55, 63, 70, 78, 101]. Some trackers are based on the generalized minimum clique problem [115]: In the work of Zamir *et al.* [70], the goal is to select a complete graph (also called *clique*) with minimal costs such that exactly one node out of each cluster is selected (see Figure 1.18). Each cluster comprises the set of detections of a particular frame. A clique represents the trajectory of one object. The algorithm is performed greedily, object by object. By simultaneously searching for K node-disjoint cliques of minimal costs (the parameter K needs to be set in advance), further improvements can be achieved [52]. Virtual nodes with pre-defined weights have to be created for objects not being detected in a frame. Due to the computational challenges of the $\mathcal{NP}$-hard problem, the method requires already computed (short) trajectories as input. These have to be generated by another MOT algorithm. False input trajectories cannot be compensated by the optimization method, thus potentially leading to error propagation.

Several state-of-the-art works employ correlation clustering [54, 63, 116] to obtain the trajectories, which is more general than the generalized minimum clique formulation. The goal is to group detections into clusters such that the agreements within the clusters (or disagreement between clusters) are maximized. Agreement and disagreement are measured in terms of all costs within the clusters or between the clusters, respectively, for which edges exist. Thus, correlation clustering allows exploiting higher-order

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

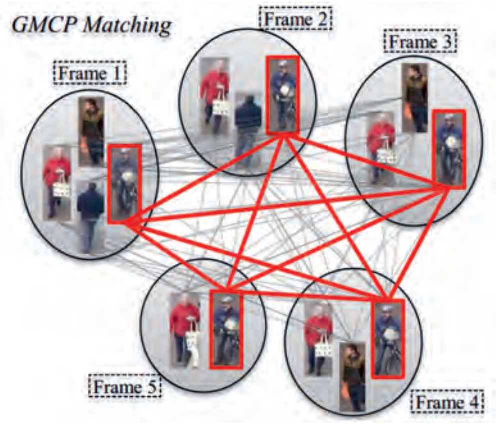


Figure 1.18: Data association graph used in the work of Zamir *et al.* [70]. The tracking problem is considered as a generalized minimum clique problem. The aim is to create a clique for each object. Missing detections are represented by virtual nodes. The costs of all clique edges are summed up in the objective function, resulting in higher-order consistencies.

consistencies, as all pairwise consistencies within a trajectory are taken into account. In contrast to the generalized minimum clique formulation, the number of clusters does not need to be known or predicted in advance. The problem is $\mathcal{NP}$-hard. Only heuristic solvers are applicable in practice [54], leading to a deteriorated result.

Connectivity priors. Data association models based on correlation clustering have been extended to include connectivity priors [64]. Corresponding solvers rely on heuristics so that the resulting assignments are generally non-optimal.

1.5 Contributions

The goal of this thesis is to propose MOT methods that enable offline computation of trajectories for all objects of a determined object class that appear in a video recording. Presented contributions focus mainly on the data association part. Scenes to which the tracking methods of this work are applied are expected to be low- to semi-crowded by the objects to be tracked.

As outlined in Section 1.3, errors occur during the detection and the feature computation step of a tracking-by-detection method. In addition, the computation of trajectories is challenging when complex data association models are employed. For this reason, existing methods rely on heuristics as briefly highlighted in Section 1.4. To overcome these limitations, this work presents two directions:

1. The strong dependence on input detections is identified as a weakness of the

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

tracking-by-detection paradigm. Therefore, MOT is formulated as a weighted graph labeling problem where all equally labeled nodes correspond to the same identity. This allows performing signal fusion, *i.e.*, augmenting object detections with additional complementary signals holistically, *e.g.*, additional detectors or sensors. Thus, advantages of two modalities are being combined. Consistency is evaluated within each and across different signals. Moreover, the data association corresponds to an implicit HO-MOT formulation. A novel solver, optimizing the underlying problem on the continuous relaxation based on the Frank-Wolfe [117] algorithm, is proposed. It produces near-optimal solutions in practice.

2. Data association can potentially compensate for errors produced by the preceding steps of a tracking-by-detection method if sufficient video information is incorporated into the decision process. This can be achieved in principle by the HO-MOT methods presented in Section 1.4. However, they commonly rely on heuristic solvers. Consequently, tracking accuracy deteriorates. Instead, a novel extension to flow networks with unary capacities is presented that incorporates connectivity priors and implements implicit HO-MOT. The problem is expressed as a *Binary Linear Program* (BLP). By studying the underlying polyhedral structure, non-trivial LP-relaxations to the optimization problem are derived and integrated into efficient separation algorithms. Finally, the BLP is solved via a cutting-plane algorithm. The resulting method is the first to provably solve the optimization problem of an HO-MOT model that incorporates long-range temporal interactions. The method delivers global optimal solutions in practice which complements existing works. Assignments are nearly optimal with respect to given input detections.

Both concepts have led to significant improvements in tracking accuracy compared to the state of the art. Details of the works are briefly elaborated in the following:

HO-MOT with Signal Fusion. The tracking-by-detection paradigm allows to significantly reduce the information of a video to be processed during the data association step, making the MOT problem tractable. However, due to errors of the detector, potentially useful information might be mistakenly discarded or outputs might be misleading in the case of false positive detections, see Section 1.3.

The first part of this work thus proposes a fusion approach to reduce the dependence on the detector by augmenting additional complementary input signals. To this end, MOT is formulated as a weighted graph labeling problem in which each input signal is assigned to an object identity by its label. Nodes and edges have weights that depend on the assigned labels. An optimal solution ensures consistent trajectories among all equally labeled input signals, thereby implementing the HO-MOT principle but extended to multiple signals. The solution is obtained from a challenging binary quadratic problem that is $\mathcal{NP}$-hard. A novel solver specifically designed for the problem is proposed. It operates on the continuous relaxation using a modification of the Frank-Wolfe algorithm [117]. A reprojection step then delivers a discrete solution. It is shown that the matrix describing the polyhedron of the graph labeling problem is totally unimodular. Consequently, each iteration step of the Frank-Wolfe algorithm and the reprojection of the continuous solution to the nearest feasible binary vector

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

can be performed efficiently by searching for the minimal entry of a vector instead of solving a linear or binary linear program. The optimal step size within the Frank-Wolfe algorithm is derived algebraically so that it can be calculated efficiently and accurately. To further improve the solver, a regularization is proposed making the method less susceptible to local minima. In addition, a hierarchical approach is presented that allows correcting approximation errors of the solver and reducing the problem size. Eventually, by iterating the process, the reduced problem can be solved globally optimal. The benefits of augmenting input detections are demonstrated in two settings, as described in the following:

(i) People detections are combined with head detections in Section 3.4. This often enables to track heavily occluded persons if their heads are visible. Thus, fusion helps to compensate for missing people detections. Also, fusion allows to reject false positive people detections that cannot be plausibly matched to head detections. Conducted experiments show quantitatively that the integration of head detections leads to significantly improved tracking results. The method, which we call *Frank-Wolfe Tracker* (FWT), won the CVPR 2017 Multi-Object Tracking Challenge against competing MOT approaches, thus demonstrating the benefit of the fusion concept. It is also shown that the modifications to the Frank-Wolfe solver are crucial in order to obtain results close to optimality.

(ii) In another setup described in Section 3.5, it is assumed that each person to be tracked is wearing one *Inertial Measurement Unit* (IMU) attached to his or her back. This minimally intrusive setting can be used, for example, in team sports when tracking players during a soccer game. The recording is supposed to be filmed by a static and calibrated camera. Conceptually, using the two input signals (video and IMU) allows resolving ambiguities caused by misleading video information, *e.g.*, in cases of similar appearances of persons (for instance in team sports) or long-term occlusions. However, incorporating the IMU signal creates new challenges, as it is highly ambiguous: The signal from an IMU at a time-step (*e.g.*, the derived orientation of a person) may fit to different persons appearing in a video recording. We term the problem of tracking people using a video signal and assigning each detection to the corresponding IMU device as *Video Inertial Multiple People Tracking* (VIMPT). We recorded a challenging dataset for VIMPT, called *VIMPT2019* [75], with a focus on non-linear motions and similar outward appearances, which are underrepresented cases in current MPT datasets.

To tackle the challenging VIMPT problem, features are presented that incorporate device-specific signals. A neural network is proposed that, when applied to a video recording, regresses a person's orientation. These estimations are linked with the orientation measurements from the IMU devices. The network utilizes a novel perspective correction that takes position of a person relative to the camera into account. This step has turned out to be crucial to obtain accurate orientation estimations. In addition, IMU acceleration measurements are compared with video-based velocities to couple motion cues. The fusion formulation ensures that computed trajectories are consistent with both the video signal and local motion information.

The proposed VIMPT method *Video Inertial Tracker* (VIT) prevents most identity switches even if the persons' appearance is very similar since the orientation regression depends only on the pose of a person and this is unrelated to outward appearance. Moreover, VIT helps in connecting detections of a person in cases where he or she has

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

been occluded for a long time-period, as the IMU signal is being transmitted independent of any occlusion. A least-squares optimization problem is proposed that robustly recovers the position of a person in cases of missing detections using both readings (video and IMU), once detections have been assigned to IMU devices. Consequently, VIT is significantly less dependent on the quality of the input detections. The coupling of video with IMU information allows to automatically label each trajectory in terms of the wearer of the IMU device, *i.e.*, an assignment of trajectories to person identities is provided, which is a valuable by-product.

Conducted experiments show significant improvements of VIT over purely video-based MPT approaches, demonstrating the effectiveness of the proposed method. The identification of persons in a video works error-free. In particular, persons are correctly identified across different sequences and when people have changed their clothing. The VIMPT problem poses an interesting extension to the MPT task with several useful advantages provided that the recording scenario at hand allows wearing minimally intrusive sensors.

We note that the gain of the fusion approach has been successfully demonstrated by the author of this dissertation in two other works that are not part of this thesis. By combining people detections with joint detections instead of head detections, the MOT method becomes even more robust to partial occlusions of persons, thus further improving tracking accuracy [101]. In another application, video information is fused with body-worn IMU sensors [118]. Here, the setup considers multiple persons (with up to two persons in the experiments of the publication) filmed by a moving camera, each of them wearing IMU devices attached to all limbs. The task is then to accurately compute the 3D pose of each person wearing the IMU devices using the information given by the camera and the IMU sensors. Solving the proposed graph labeling formulation provides trajectories that are consistent with the video information and the IMU sensors. In particular, all persons in a video wearing the IMU sensors are automatically identified which makes it possible to estimate and correct heading drift and estimate the relative positions between IMU-equipped persons. The combination of the two sensor modalities thus results in very accurate human pose estimations. Most importantly, by using body-worn IMUs and a single, non-static, hand-held camera, the method allows, for the first time, to accurately capture poses of multiple humans during daily activities, *e.g.*, walking around a city.

Summary of contributions:

- Extension of the tracking-by-detection approach by signal fusion resulting in an HO-MOT model that is less dependent on input detections.
- New solver for the underlying weighted graph labeling problem with nearly optimal solutions.
- Fusing people detections with head detections via FWT. The fusion approach won the CVPR 2017 Multi-Object Tracking Challenge.
- VIMPT2019, a dataset for VIMPT with a focus on scenes that are challenging for commonly employed pairwise features.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

- An accurate VIMPT method (VIT) using perspective-aware orientation regression and acceleration measurements. It provides (in the experiments error-free) labeling of trajectories without additional computational costs. VIT is significantly less dependent on the video signal, input detections, appearance and motion models when compared with purely video-based methods.
- Missing detections are robustly reconstructed in VIT by fusing video signals with IMU information within a non-linear optimization problem.

Lifted Disjoint Paths. The second part of this work focuses on integrating long-range temporal interactions into the data association step without relying on heuristic solvers. Conceptually, this allows for the compensation of errors generated by the detector and to be less susceptible to misleading pairwise costs.

The min-cost network flow problem constraint to unary capacities, also called *disjoint paths problem*, serves as a natural formulation for trajectories within MOT [61]. However, long-term information can only be incorporated to a limited extent, see Figure 1.8. This work proposes a novel extension, the *Lifted Disjoint Paths* (LDP) problem, by augmenting the network flow graph with additional lifted edges to provide path connectivity priors. Lifted edges facilitate the re-identification of persons, help to prevent ID switches, and enable HO-MOT. Since the disjoint paths problem has many applications in discrete optimization [119], the extension to LDP may be relevant to other fields as well.

The new formulation is analyzed in terms of its complexity class. It is proven that the problem is $\mathcal{NP}$-hard by reduction[7] from integer multi-commodity flow and 3-SAT. On the one hand, this makes the problem interesting for further research in computational complexity theory. On the other hand, the $\mathcal{NP}$-hard property makes the problem difficult.

A formulation of the LDP problem as a binary linear program is derived. By further studying the polyhedral structure, improved LP-relaxations are proposed that are tighter than using the linear constraints of the initial LDP formulation. To handle the exponentially many linear constraints, efficient separation algorithms are proposed and embedded into a cutting-plane procedure. The resulting solver enables global optimal solutions in practice. Existing methods either employ heuristics on complex models or global optimization on simplified models so that they do not take full advantage of the provided information, especially long-range interactions. In contrast, this work presents, to our knowledge, the first method to solve HO-MOT problems via global optimization.

Robust pairwise features combining visual and motion cues are presented, suitable for relating detections over long time periods. The complete setup takes pairwise interactions between detections into account that are at most 60 frames apart. It is further limited to a temporal distance of 2sec. Experiments reveal that the proposed method delivers nearly optimal assignments with respect to given input detections. Extensive experiments also show that the method outperforms the state of the art in MOT by a large margin.

Summary of contributions:

---

[7] The terminology is introduced in Section 2.5.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

- Novel extension of the disjoint paths problem such that it includes connectivity priors. Proofs are provided that the new formulation is $\mathcal{NP}$-hard.
- Improved linear inequalities with tighter LP-relaxations than a direct formulation as a binary linear program.
- Efficient separation algorithms to obtain global optimal solutions in practice.
- First global optimal solver for implicit HO-MOT with long-range temporal interactions.
- Robust features to take full advantage of long-term information.

1.6 List of Publications

During the course of this thesis, several peer-reviewed publications have been published by the author of this dissertation at computer vision conferences and journals, which are presented below. The first five listed publications cover the topic of HO-MOT using signal fusion (Chapter 3) by incorporating additional detectors and sensors. The subsequent sixth publication builds the basis for Chapter 4. The seventh and eighth publications present an approximate solver for the LDP problem of Chapter 4. Finally, two additional works cover MOT using a novel trajectory model based on a minimum cost arborescene problem that leads to efficient computations and a robust method.

Two methods were awarded a prize. The fusion of head and people detections [55] won the CVPR 2017 Multi-Object Tracking Challenge. The method using minimum cost arborescene [120] won the second place on the WACV 2015 Challenge.

[55] **Roberto Henschel**, Laura Leal-Taixé, Daniel Cremers, Bodo Rosenhahn: "Fusion of Head and Full-Body Detectors for Multi-Object Tracking". *In: CVPR Workshop on Joint Detection, Tracking, and Prediction in the Wild (CVPRW), 2018.*

Abstract: In order to track all persons in a scene, the tracking-by-detection paradigm has proven to be a very effective approach. Yet, relying solely on a single detector is also a major limitation, as useful image information might be ignored. Consequently, this work demonstrates how to fuse two detectors for a tracking method. To obtain the trajectories, we propose to formulate tracking as a weighted graph labeling problem, resulting in a binary quadratic program. As such problems are $\mathcal{NP}$-hard, the solution can only be approximated. Based on the Frank-Wolfe algorithm, we present a new solver that is crucial to handle such difficult problems. Evaluation on pedestrian tracking is provided for multiple scenarios, showing superior results over single detector tracking and standard QP-solvers. Finally, our tracker ranks 2nd on the MOT16 benchmark and 1st on the new MOT17 benchmark, outperforming over 90 trackers.

[101] **Roberto Henschel**, Yunzhe Zou, Bodo Rosenhahn: "Multiple People Tracking using Body and Joint Detections". *In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.*

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Abstract: Most multiple people tracking methods compute trajectories based on the tracking-by-detection paradigm. Consequently, the performance depends to a large extent on the quality of the employed input detections. However, despite enormous progress in recent years, partially occluded people are still often not recognized. Also, many correct detections are mistakenly discarded when the non-maximum suppression is performed. Improving the tracking performance thus requires augmenting the coarse input. Well-suited for this task are fine-graded body joint detections, as they allow to locate even strongly occluded persons. Thus in this work, we analyze the suitability of including joint detections for multiple people tracking. We introduce different affinities between the two detection types and evaluate their performances. Tracking is then performed within a near-online framework based on a min-cost graph labeling formulation. As a result, our framework can recover heavily occluded persons and solve the data association efficiently. We evaluate our framework on the MOT17 benchmark. Experimental results demonstrate that our framework achieves state-of-the-art results.

[121] **Roberto Henschel**, Timo von Marcard, Bodo Rosenhahn: “Simultaneous Identification and Tracking of Multiple People using Video and IMUs”. *In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.*

Abstract: Most modern approaches for multiple people tracking rely on human appearance to exploit similarity between person detections. In this work, we propose an alternative tracking method that does not depend on visual appearance and is still capable of dealing with very dynamic motions and long-term occlusions. We make this feasible by: (i) incorporating additional information from body-worn inertial sensors, (ii) designing a neural network to relate person detections to orientation measurements, and (iii) formulating a graph labeling problem to obtain a tracking solution that is globally consistent with the video and inertial recordings. We evaluate our approach on several challenging tracking sequences and achieve a very high IDF1 score of 91.2%. We outperform appearance-based baselines in scenarios where appearance is less informative and are on-par in situations with discriminative people appearance.

[122] **Roberto Henschel**, Timo von Marcard, Bodo Rosenhahn: “Accurate Long-Term Multiple People Tracking using Video and Body-Worn IMUs”. *In: Transactions on Image Processing (TIP), 2020.*

Abstract: Most modern approaches for video-based multiple people tracking rely on human appearance to exploit similarities between person detections. Consequently, tracking accuracy degrades if this kind of information is not discriminative or if people change apparel. In contrast, we present a method to fuse video information with additional motion signals from body-worn IMUs. In particular, we propose a neural network to relate person detections with IMU orientations, and formulate a graph labeling problem to obtain a tracking solution that is globally consistent with the video and inertial recordings. The fusion of visual and inertial cues provides several advantages. The association of detection boxes in the video and IMU devices is based on motion, which is independent of a person’s outward appearance. Furthermore, inertial sensors provide motion information irrespective

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

of visual occlusions. Hence, once detections in the video are associated with an IMU device, intermediate positions can be reconstructed from corresponding inertial sensor data, which would be unstable using video only. We release a dataset of challenging tracking sequences containing video and IMU recordings together with ground truth annotations. We evaluate our approach on our new dataset, achieving an average IDF1 score of 91.2%. The proposed method is applicable to any situation that allows one to equip people with inertial sensors.

[118] Timo von Marcard, **Roberto Henschel**, Michael J. Black, Bodo Rosenhahn, Gerard Pons-Moll: "Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera". *In: Proceedings of the European Conference on Computer Vision (ECCV), 2018.*

Abstract: In this work, we propose a method that combines a single hand-held camera and a set of IMUs attached at the body limbs to estimate accurate 3D poses in the wild. This poses many new challenges: the moving camera, heading drift, cluttered background, occlusions, and many people visible in the video. We associate 2D pose detections in each image to the corresponding IMU-equipped persons by solving a novel graph based optimization problem that forces 3D to 2D coherency within a frame and across long range frames. Given associations, we jointly optimize the pose of a statistical body model, the camera pose and heading drift using a continuous optimization framework. We validated our method on the TotalCapture dataset, which provides video and IMU synchronized with ground truth. We obtain an accuracy of 26mm, which makes it accurate enough to serve as a benchmark for image-based 3D pose estimation in the wild. Using our method, we recorded 3D Poses in the Wild (3DPW), a new dataset consisting of more than 51,000 frames with accurate 3D pose in challenging sequences, including walking in the city, going up-stairs, having coffee or taking the bus. We make the reconstructed 3D poses, video, IMU and 3D models available for research purposes at `http://virtualhumans.mpi-inf.mpg.de/3DPW`.

[50] Andrea Hornakova[8], **Roberto Henschel**[8], Bodo Rosenhahn, Paul Swoboda: "Lifted Disjoint Paths with Application in Multiple Object Tracking". *In: International Conference on Machine Learning (ICML), 2020.*

Abstract: We present an extension to the disjoint paths problem in which additional lifted edges are introduced to provide path connectivity priors. We call the resulting optimization problem the lifted disjoint paths problem. We show that this problem is $\mathcal{NP}$-hard by reduction from integer multi commodity flow and 3-SAT. To enable practical global optimization, we propose several classes of linear inequalities that produce a high-quality LP-relaxation. Additionally, we propose efficient cutting plane algorithms for separating the proposed linear inequalities. The lifted disjoint paths problem is a natural model for multiple object tracking and allows an elegant mathematical formulation for long-range temporal interactions. Lifted edges help to prevent ID switches and to re-identify persons. Our lifted disjoint paths tracker leads on all three main benchmarks of the MOT challenge, improving significantly over state of the art.

---

[8] Shared first authorship.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[123] Andrea Hornakova, Timo Kaiser, Bodo Rosenhahn, Paul Swoboda, **Roberto Henschel**: “Higher Order Multiple Object Tracking for Crowded Scenes”. *In: IEEE Conference on Computer Vision and Pattern Recognition (CVPRW), 2021.*

Abstract: The lifted disjoint paths formulation is a natural model for multiple object tracking. This model is able to obtain state-of-the-art results but is $\mathcal{NP}$-hard. We present an efficient approximate message passing solver for LDP and integrate it into a multiple object tracker, which scales to very large instances that come from long and crowded scenes. We achieve comparable or better performance than state-of-the-art methods on MOT15/16/17 benchmarks and comparable results on the MOT20 benchmark. This has been out of reach up to now for known LDP-solvers due to the problem size and complexity of MOT20.

[124] Andrea Hornakova, Timo Kaiser, Paul Swoboda, Michal Rolinek, Bodo Rosenhahn, **Roberto Henschel**: “Making Higher Order MOT Scalable: An Efficient Approximate Solver for Lifted Disjoint Paths”. *In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021.*

Abstract: We present an efficient approximate message passing solver for the lifted disjoint paths problem (LDP), a natural but NP-hard model for multiple object tracking (MOT). Our tracker scales to very large instances that come from long and crowded MOT sequences. Our approximate solver enables us to process the MOT15/16/17 benchmarks without sacrificing solution quality and allows for solving MOT20, which has been out of reach up to now for LDP solvers due to its size and complexity. On all these four standard MOT benchmarks we achieve performance comparable or better than current state-of-the-art methods including a tracker based on an optimal LDP solver.

[120] **Roberto Henschel**, Laura Leal-Taixé, Bodo Rosenhahn : “Solving Multiple People Tracking In A Minimum Cost Arborescence”. *In: Winter Conference on Applications of Computer Vision Workshops (WACVW), 2015.*

Abstract: For many applications of computer vision, it is necessary to localize and track humans that appear in a video sequence. Multiple people tracking has thus evolved as an ongoing research topic in the computer vision domain. A commonly used approach to solve the data association problem within the tracking task is to apply a hierarchical tracklet framework. Although there has been great progress in such a model, mainly due to its good bootstrapping capabilities, so far little attention has been drawn to improve the quality of the tracklets themselves. A main issue of the hierarchical frameworks, as used in the common literature, is that they make hard decisions at each iteration of the association step. Especially in ambiguous situations, tracklets are still being merged or removed so that the method is prone to error propagation. To avoid these problems, we propose a new framework that prevents unreliable decisions. Instead, unclear aggregations are being postponed to a later iteration when more information is available. To maintain the possible associations of tracklets under challenging situations, we propose a new trajectory model, which we call *tree tracklets*. While recent multiple people trackers model the association problem mainly in a flow network, we employ a rooted, directed and weighted graph, which is of a simpler structure,

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

in particular has fewer nodes and edges. Thereby, we obtain the global optimal solution of each iteration in linear time in the number of nodes by computing a minimum cost arborescence.

[106] **Roberto Henschel**, Laura Leal-Taixé, Bodo Rosenhahn: “Efficient Multiple People Tracking Using Minimum Cost Arborescences”. *In: German Conference on Pattern Recognition (GCPR), 2014.*

Abstract: We present a new global optimization approach for multiple people tracking based on a hierarchical tracklet framework. A new type of tracklets is introduced, which we call *tree tracklets*. They contain bifurcations to naturally deal with ambiguous tracking situations. Difficult decisions are postponed to a later iteration of the hierarchical framework when more information is available. We cast the optimization problem as a minimum cost arborescence problem in an acyclic directed graph, where a tracking solution can be obtained in linear time. Experiments on six publicly available datasets show that the method performs well when compared to state-of-the art tracking algorithms.

In addition, a dataset has been recorded and published for research purposes.

[75] **Roberto Henschel**, Timo von Marcard, Bodo Rosenhahn. “VIMPT2019 - Video Inertial Multiple People Tracking Dataset”. `https://www.tnt.uni-hannover.de/de/project/VIMPT2019/`, 2019.

VIMPT2019 is a dataset for multiple people tracking that is recorded by a camera, providing also the motion information from a body-worn IMU for each visible person. The dataset consists of 6 soccer sequences that are challenging due to frequent occlusions, appearance ambiguities (because of the worn uniforms), and non-linear motions. In addition, an outdoor sequence filmed in a park environment, which is similar to the typical multiple person tracking sequence, is provided. In each sequence, 8 persons wear an IMU sensor that is synchronized to a calibrated camera.

1.7 Outline

The structure of the thesis is presented, followed by a graphical overview in Figure 1.19.

Chapter 1 (Introduction): Introduces the reader to the problem statement of multiple object tracking, summarizes the novelties of this work, and compares it to related work.

Chapter 2 (Fundamentals): Defines the basic terminologies used throughout this work and provides the fundamentals necessary to develop and assess the proposed tracking methods. It covers graph theory, machine learning, complexity theory, and optimization theory. Regarding MOT, employed object detectors, datasets, pairwise features, and metrics are introduced.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Chapter 3 (HO-MOT with Signal Fusion): Presents an MOT formulation to augment tracking-by-detection by additional signals, resulting in an HO-MOT model so that computed trajectories are consistent with respect to all provided data. A novel approximate solver is proposed to solve the underlying optimization problem. It leads to high-quality solutions and is considerably faster than an optimal solver. The benefit of the entire framework is shown in different settings. For video-based multiple people tracking, fusing two types of input detections (people detections together with head detections) with the method leads to significantly improved tracking results. In addition, the framework is used to fuse video information with signals from body-worn IMUs. The experiments show that such a fusion leads to a tracking method that is less dependent on the detection quality and on motion and appearance models of persons. The method outperforms purely video-based MOT methods on sequences with ambiguous appearance information by a large margin. The method tracks people and simultaneously identifies them (even across sequences and when clothes have been changed). The chapter is based on previously published work [55, 121, 122].

Chapter 4 (Lifted Disjoint Paths): Presents a novel tracking formulation to perform video-based HO-MOT, enabling to incorporate connectivity priors for long-range temporal interactions. The formulation is shown to be $\mathcal{NP}$-hard, making it challenging to solve. Yet, a global optimal solver is developed that delivers fast solutions in practice. The method represents, to our knowledge, the first work to solve HO-MOT incorporating long-range interactions without heuristics. Long-term connections up to 60 frames apart are exploited during the decision process. Extensive experiments on several multi-person tracking datasets show that the method leads to very high tracking accuracies, outperforming the state of the art by a large margin. Assignments are nearly optimal with respect to provided input detections. The chapter is based on previously published work [50].

Chapter 5 (Conclusions): Summarizes contributions and results of the methods of Chapter 3 and Chapter 4. In addition, their limitations are discussed and addressed in an outlook for future work.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

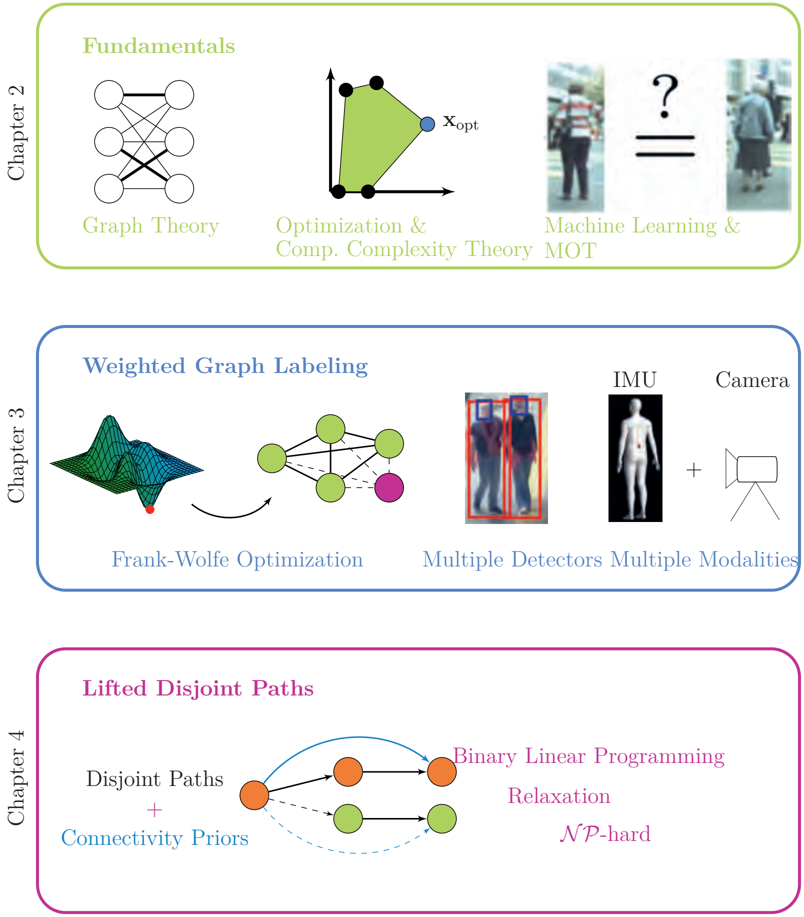


Figure 1.19: Thesis overview.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

2 Fundamentals

The MOT methods presented in this work are based on different fields of mathematics and computer vision, especially discrete optimization, machine learning, and object detection. This chapter briefly introduces the necessary basics and defines notation and terminology used throughout this work.

2.1 Sets, Maps, and Matrices

We settle the basic terminologies used for sets, maps, and matrices.

Sets. We denote by $\mathbb{N}, \mathbb{Z}$, and $\mathbb{R}$ the natural numbers (excluding zero), integer numbers, and real numbers, respectively. By $\mathbb{Z}_{\geq 0}$ and $\mathbb{R}_{\geq 0}$, we denote the non-negative integers and non-negative real numbers, respectively. For $n_1, n_2 \in \mathbb{Z}$, we define

$$[n_1 : n_2] \coloneqq \{t \in \mathbb{Z} \mid n_1 \leq t \leq n_2\}, \quad [n_2] \coloneqq [1 : n_2]\,, \quad [n_2]_0 \coloneqq [0 : n_2]\,. \tag{2.1}$$

For set A, we define the set of two-element sets $A^{(2)} \coloneqq \{\{x, y\} \mid x, y \in A, x \neq y\}$. For set $A \subseteq \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^n$, we define the *Minkowski sum* as $A + \mathbf{x} \coloneqq \{\mathbf{a} + \mathbf{x} \mid \mathbf{a} \in A\}$.

A set $X \neq \emptyset$ is *partitioned* by sets $X_1, \ldots, X_n \subseteq X$ if $X = \bigcup_{k=1}^{n} X_i$ and $X_i \cap X_j = \emptyset$ for all $i \neq j$. In this case, we write $X = X_1 \sqcup \ldots \sqcup X_n$. For sets A, B, we set $A \subsetneq B$ if and only if $A \subset B$ and $A \neq B$.

Maps. For a map $f : X \to Y$, we may write $f \in Y^X$. For a finite set X with $n = |X|$, we identify a map $f \in \mathbb{R}^X$ with the vector $\mathbf{f} \coloneqq (f(x))_{x \in X} \in \mathbb{R}^n$. For set X, $n \in \mathbb{N}$, and $i \in [n]$, we define $\pi_i : X^n \to X$ as the projection on the i-th component so that $\pi_i(\mathbf{x}) = x_i$ for all $\mathbf{x} = (x_1, \ldots, x_n) \in X^n$.

Matrices. We frequently use the vector space identification $\mathbb{R}^{n \times m} \cong \mathbb{R}^{nm}$ for $n, m \in \mathbb{N}$ without additional notation. For a matrix $\mathbf{X} = (x_{r,c})_{r \in [n], c \in [m]} \in \mathbb{R}^{n \times m}$, $j \in [n]$ and $k \in [m]$, we define the row vector $\mathbf{X}_{[j,:]}$ and column vector $\mathbf{X}_{[:,k]}$ by

$$\mathbf{X}_{[j,:]} \coloneqq (x_{j,c})_{c \in [m]} \in \mathbb{R}^{1 \times m}, \quad \mathbf{X}_{[:,k]} \coloneqq (x_{r,k})_{r \in [n]} \in \mathbb{R}^{n \times 1}\,, \tag{2.2}$$

respectively. For a tensor $\mathbf{X} \in \mathbb{R}^{n \times m \times p}$ and $u \in [p]$, we define $\mathbf{X}_{[:,:,u]} \in \mathbb{R}^{n \times m}$ accordingly. Throughout this work, we denote by $\mathbf{0}$ and $\mathbf{1}$ the zero matrix and matrix of ones, respectively. If it is clear from the context, we omit the dimension for $\mathbf{0}$ and $\mathbf{1}$.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

2.2 Probability Theory

In order to formulate the tracking task as obtaining the most likely trajectories and to incorporate measurement uncertainties, probability theory is employed.

We provide the basic terminologies and refer to Georgii [69] for further details.

Definition 2.1. Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space with sample space Ω, σ-algebra $\mathcal{A}$ and probability measure P. Let $(\Omega', \mathcal{A}')$ be a measurable space. A mapping $X : \Omega \to \Omega'$ is a *random variable* if $X^{-1}(A) \in \mathcal{A}$ for all $A \in \mathcal{A}'$. For random variable X and $\omega \in \Omega$, $x = X(\omega)$ is a *realization*, which is an observed value of X.

In this work, we will only consider random variables $X : \Omega \to \Omega'$ with $\Omega' \subseteq \mathbb{R}$ and finite $X(\Omega)$. Such a mapping is called a *discrete random variable*.

By abuse of notation, we may omit random variables if the statement remains clear from the context. Thus for random variable $X : \Omega \to \mathbb{R}$ and $x \in \mathbb{R}$, we may write $\mathsf{P}(x) := \mathsf{P}(X = x) := \mathsf{P}(\{w \in \Omega \mid X(\omega) = x\})$.

Definition 2.2. The *expected value* of a discrete random variable $X : \Omega \to \mathbb{R}$ is given by

$$\mathbb{E}[X] := \sum_{x \in X(\Omega)} x \mathsf{P}(X = x)\,. \tag{2.3}$$

2.3 Graph Theory

The tracking methods of Chapter 3 and Chapter 4 formulate the MOT task in terms of a *graph*, a mathematical construct to represent entities (*e.g.*, detections) and to model pairwise relations between them (*e.g.*, associations of detections to a person).

A very brief introduction to the basic concepts of graph theory is presented, based on Jungnickel [125].

Definition 2.3. A *(simple) undirected graph* is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of sets with $|\mathcal{V}| < \infty$ and $\mathcal{E} \subseteq \mathcal{V}^{(2)}$. An element $v \in \mathcal{V}$ is a *vertex* or *node*. Each element $e \in \mathcal{E}$ is an *edge*. Further definitions characterize nodes and their relations to edges:

- Vertices $v, u \in \mathcal{V}$ are *adjacent* if $\{v, u\} \in \mathcal{E}$.
- A vertex $v \in \mathcal{V}$ is *incident* to an edge $e \in \mathcal{E}$ if $v \in e$.
- A *labeling* of a graph $\mathcal{G}$ is a map $\mathcal{L} : \mathcal{V} \to [n]$ for $n \in \mathbb{N}$, which defines for each node $v \in \mathcal{V}$ its associated *label* $\mathcal{L}(v)$. A graph is *vertex-labeled* if a bijection $\mathcal{L} : \mathcal{V} \to [\,|\mathcal{V}|\,]$ is given.
- The *neighborhood* $N_{\mathcal{G}}(v)$ of a vertex $v \in \mathcal{V}$ is the set of vertices adjacent to v. The value $d_G(v) := |N_{\mathcal{G}}(v)|$ is the *degree* of v.

For an edge $e = \{v, u\}$, we shall simply write $e = vu$ (or $e = uv$). We may refer to the set of nodes and edges of $\mathcal{G}$ as $\mathcal{G}_{\mathcal{V}}$ and $\mathcal{G}_{\mathcal{E}}$, respectively. An example graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{u, v\}$ and $\mathcal{E} = \{\{v, u\}\}$ is drawn in Figure 2.1(a). Each node of $\mathcal{G}$ has degree 1.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

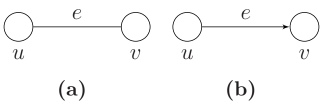


Figure 2.1: (a) A graph consisting of two nodes (u and v) and one edge (e). We represent a node by a circle and an edge by a line. Note that a graph does not specify the positions of the nodes. (b) A directed path. Here, the edge $e = uv$ is drawn.

Remark 2.4. All graphs used in this work are assumed to be vertex-labeled without explicitly stating it. In particular, this defines an ordering on the vertices so that we can identify a map $x \in \mathbb{R}^{\mathcal{V}}$ with a vector $\mathbf{x} \in \mathbb{R}^{|\mathcal{V}|}$.

An important task studied in graph theory is to select edges of a given graph according to an optimality criterion. This is used in the tracking methods proposed in Chapter 3 and Chapter 4 in which each edge represents an association of two detections to the same object. Also, during the evaluation of a tracking method, edges are selected in order to link detections with ground truth data (see Section 2.7.4). For the evaluation, any node must be incident to at most one selected edge, which is formalized using *matchings* as follows:

Definition 2.5. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph and $M \subseteq \mathcal{E}$.

- $M \subseteq \mathcal{E}$ is called *matching* if any two distinct edges are node-disjoint:
$$\forall e, f \in M : e \neq f \Longrightarrow e \cap f = \emptyset \,. \tag{2.4}$$
- For matching M and edge $uv \in M$, the vertices u and v are *matched.*
- M is a *maximum* matching if (i) M is a matching and (ii) $|M| \geq |M'|$ for all matchings $M' \subseteq \mathcal{E}$.

An example of a maximum matching is given in Figure 2.2. Matchings can be computed efficiently using the Hungarian algorithm (see Section 2.3.2).

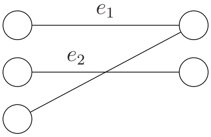


Figure 2.2: A maximum matching of the depicted graph is given by $M = \{e_1, e_2\}$.

A graph can describe the set of possible trajectories and their probabilities by assigning values to nodes or edges.

Definition 2.6. Given a graph $(\mathcal{V}, \mathcal{E})$ and let $\mathbf{w} \in \mathbb{R}^{|\mathcal{E}|}$ and $\mathbf{c} \in \mathbb{R}^{|\mathcal{V}|}$ be given. The tuple $(\mathcal{V}, \mathcal{E}, \mathbf{w})$ is an *edge-weighted* graph. Accordingly, $(\mathcal{V}, \mathcal{E}, \mathbf{c})$ is a *vertex-weighted* graph.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Remark 2.7. Edge weights may also be provided equivalently in terms of a symmetric matrix $\mathbf{Q} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$. Consequently, all diagonal entries of $\mathbf{Q}$ must be zero.

Chapter 3 formulates the MOT task using a vertex-weighted graph with a matrix $\mathbf{Q}$ as described in Remark 2.7. The tracking method presented in Chapter 4 uses a vertex- and edge-weighted graph.

For each edge $e = vu$, a direction ((v, u) or (u, v)) can be defined, see also Figure 2.1(b). This is used in Chapter 4 to formulate the MOT problem, as it enables to naturally model the temporal information of a video recording.

Definition 2.8. A *(simple) directed graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set $\mathcal{V}$ of nodes and a set $\mathcal{E} \subseteq (\mathcal{V} \times \mathcal{V}) \setminus \{(v, v) \mid v \in \mathcal{V}\}$ of edges. In a directed graph, an edge $e = (v, u) =: vu$ is an *incoming* edge for vertex u, while it is an *outgoing* edge for vertex v. The set of incoming and outgoing vertices for node v is denoted as $N^-_{\mathcal{G}}(v)$ and $N^+_{\mathcal{G}}(v)$, respectively. Accordingly, $d^-_{\mathcal{G}}(v) := |N^-_{\mathcal{G}}(v)|$ and $d^+_{\mathcal{G}}(v) := |N^+_{\mathcal{G}}(v)|$ denote the number of incoming and outgoing edges at node $v \in \mathcal{V}$, respectively.

2.3.1 Important graph classes

Several graph classes are introduced that will be used throughout this work.

Definition 2.9. A (directed) *subgraph* $\mathcal{G}'$ of a (directed) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, denoted by $\mathcal{G}' \subseteq \mathcal{G}$, is defined to be a (directed) graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ with $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{E}' \subseteq \mathcal{E}$. As a special case, given a subset of nodes $\mathcal{V}' \subseteq \mathcal{V}$, we define the *induced* subgraph $\mathcal{G}[\mathcal{V}'] := (\mathcal{V}', \mathcal{E}')$, where the edge set is given via $\mathcal{E}' := \{vu \in \mathcal{E} \mid v, u \in \mathcal{V}'\}$.

Induced subgraphs keep all edges that have both endpoints in the defined vertex set. The hierarchical solver proposed in Section 3.3.2.4 uses induced subgraphs to select for each computed trajectory the corresponding subgraph, which enables correcting wrong assignments.

Definition 2.10. A (directed) *path* is a non-empty (directed) graph $P = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \{v_1, \ldots, v_k\}$ and edges $\mathcal{E} = \{v_i v_{i+1} \mid i \in [k-1]\}$, while all vertices v_i have to be distinct. Two paths $P_1, P_2 \subseteq \mathcal{G}$ are *disjoint* if they do not share a common vertex.

The MOT formulation of Chapter 4 represents each trajectory by a path. Also, paths allow characterizing graph connectivity.

Definition 2.11. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a (directed) graph.

- Nodes $v, u \in \mathcal{V}$ are *connected* if a (directed) path P in $\mathcal{G}$ exists with $v, u \in P_{\mathcal{V}}$.
- An edge $e = vu \in \mathcal{E}$ is a *bridge* if v, u are not connected in $(\mathcal{V}, \mathcal{E} \setminus \{e\})$.

For instance, any edge of the graph depicted in Figure 2.3(a) is a bridge, while no edge of the graph depicted in Figure 2.3(b) is a bridge.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

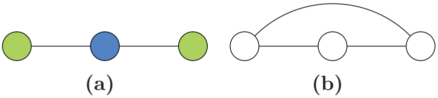


Figure 2.3: (a) The graph is a path. The node coloring shows that the graph is also bipartite. (b) The graph is a cycle and also a complete graph.

Definition 2.12. A (directed) *cycle* is a non-empty (directed) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $\mathcal{V} = \{v_1, \ldots, v_k\}$ and edges $\mathcal{E} = \{v_i v_{i+1} \mid i \in [k-1]\} \cup \{v_1 v_k\}$, while all vertices v_i have to be distinct. A (directed) graph $\mathcal{G}$ is *acyclic* if no subgraph of $\mathcal{G}$ is a cycle.

Definition 2.13. In a *complete* graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, any two vertices of $\mathcal{V}$ are connected by an edge.

Assignments of identities to detections are computed using complete graphs in Chapter 3 so that all connections of a trajectory γ, *i.e.*, all edges of the subgraph induced by the detections of γ, are taken into account.

A key role in computing matchings play bipartite graphs, which are defined in the following.

Definition 2.14. An undirected graph $\mathcal{G} = (\mathcal{V}_1 \sqcup \mathcal{V}_2, \mathcal{E})$ is *bipartite* if $\{v, u\} \not\subset \mathcal{V}_1 \wedge \{v, u\} \not\subset \mathcal{V}_2$ for all $\{v, u\} \in \mathcal{E}$. $\mathcal{G}$ is a *complete* bipartite graph if for all $v \in \mathcal{V}_1$ and for all $u \in \mathcal{V}_2$, $vu \in \mathcal{E}$. If $|\mathcal{V}_1| = n = |\mathcal{V}_2|$, we shall write $\mathcal{G} = K_{n,n}$.

Examples of the above graph classes are shown in Figure 2.3.

2.3.2 Computations on graphs

Bridges. Given a graph $\mathcal{G}$, the shortest distance d between two nodes $v, u \in \mathcal{V}$ can be computed using a function `Shortest-Distance`$(v, u, \mathcal{G})$ that returns $d = \infty$ if v and u are not connected in $\mathcal{G}$. Otherwise, it returns the length of a shortest path P that connects v with u, where the length of P is defined as the number of edges $|P_{\mathcal{E}}|$. The Dijkstra algorithm [126] is a suitable and efficient shortest path implementation with time-complexity $\mathcal{O}(|\mathcal{V}| + |\mathcal{E}|)$. This allows detecting bridges in a graph in $\mathcal{O}(|\mathcal{E}|(|\mathcal{V}| + |\mathcal{E}|))$, which is used in Section 3.3.2.4 to improve the objective value initially returned by the proposed solver. Pseudo-code of a bridge-detection algorithm is presented in Algorithm 2.1.

Matchings. To compute matchings in a graph, we consider an edge-weighted bipartite graph $\mathcal{G} = (\mathcal{V}_1 \sqcup \mathcal{V}_2, \mathcal{E}, \mathbf{w})$. The *weight* of a matching $M \subseteq \mathcal{E}$ is defined as $w(M) := \sum_{e \in M} w_e$. A matching M is a *maximum-weight* matching if $w(M) \geq w(M')$ for all matchings $M' \subseteq \mathcal{E}$ of $\mathcal{G}$. A matching M is *perfect* if each vertex $v \in \mathcal{V}$ is incident to a matching edge $e \in M$. A maximum and perfect matching of a complete bipartite graph $K_{n,n}$ is called *optimal* matching and can be computed using the Hungarian matching algorithm.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Algorithm 2.1: Bridge detection

```
Input   : Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.
Output  : Bridge set $\mathcal{E}_{\text{bridge}}$.

$\mathcal{E}_{\text{bridge}} = \emptyset$
for $e = vu \in \mathcal{E}$ do
    $\mathcal{G}' \leftarrow (\mathcal{V}, \mathcal{E} \setminus \{e\})$
    $d \leftarrow$ Shortest-Distance$(v, u, \mathcal{G}')$
    if $d = \infty$ then
        $\mathcal{E}_{\text{bridge}} \leftarrow \mathcal{E}_{\text{bridge}} \cup \{e\}$
    end if
end for
```

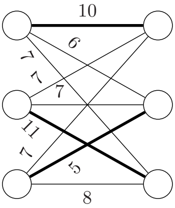


Figure 2.4: The optimal matching of the graph has weight 26, which is composed of the thick edges.

Theorem 2.15 (Hungarian matching)**.** *For a complete bipartite graph $K_{n,n}$ with non-negative edge weights, an optimal matching can be obtained in $\mathcal{O}(n^3)$.*

Proof. See Kuhn [100]. □

An example of an optimal matching is shown in Figure 2.4. The Hungarian matching algorithm is used to establish correspondences between computed and ground truth trajectories to evaluate tracking results (see Section 2.7.4).

Remark 2.16. The problem of finding a maximum-weight matching can be performed using a canonical form without loss of generality: We can assume that (i) $\mathbf{w}$ is non-negative, (ii) $|\mathcal{V}_1| = |\mathcal{V}_2|$, and (iii) $\mathcal{G}$ is a complete graph. To ensure (i), a weight vector $\mathbf{w}$ with negative components can be replaced by $\mathbf{w} - c$ with $c = \min_{e \in \mathcal{E}} w_e$. To ensure (ii), we can add nodes with only zero-weighted incident edges. To ensure (iii), we can add missing edges with zero weight.

2.4 Machine Learning

Machine learning is a field of artificial intelligence that is frequently used in this thesis, *e.g.*, to detect persons within an image and to find meaningful pairwise features between two detection boxes.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

A commonly accepted definition is that a computer program is said to *learn* from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [127]. Machine learning allows solving certain tasks without explicitly programming an algorithm that directly solves the problem. Instead, the information contained in a dataset is used to automatically infer the correct adjustment of an algorithm for the desired task.

2.4.1 Supervised learning

A subfield of machine learning is *supervised learning*, whose aim is to approximate or to *learn* a map based on a given set of input-output pairs. A very brief introduction to the topic is provided, following Mohri *et al.* [128].

Let $\underline{f} : X \to Y$ be a mapping that is not explicitly given, *i.e.*, it is not known how to define a map that returns $\underline{f}(x)$ for all $x \in X$. Yet, for a finite subset $S \subsetneq X$ the input-output relation is provided, *i.e.*, we are given the restriction $\underline{f}|_S$ of $\underline{f}$ to S. The set $\{(s, \underline{f}(s)) \mid s \in S\} \subset X \times Y$ is a *training set* using *labeled samples* S.

For example, X could be a set of all possible images that can be taken from a camera showing either a cat or a dog, $Y = \{0, 1\}$, and $\underline{f}(x)$ outputs 0 if $x \in X$ shows a dog, and 1 otherwise. In this case, $S \subsetneq X$ is a finite collection of images of cats and dogs, where the correct output is defined by human supervision for each input.

The set X contains all possible *examples* and the set Y comprises all feasible *labels*. To learn $\underline{f}$ from $\underline{f}|_S$, a hypothesis set $\mathcal{H}$ of maps from X to Y is considered. Finally, the task of supervised learning is to find the best hypothesis $h \in \mathcal{H}$ being as similar to $\underline{f}$ as possible. Note that often, $\underline{f} \notin \mathcal{H}$. As only $\underline{f}|_S$ is given and no further constraints on $\underline{f}$ are imposed, statistical methods are used to infer an approximation of $\underline{f}$.

Given a probability distribution P on X, the task of supervised learning is to minimize the expected prediction error when using a map $h \in \mathcal{H}$ instead of $\underline{f}$. To measure how much a prediction deviates from the expected output, a *loss* function $L : Y \times Y \to \mathbb{R}_{\geq 0}$ is used. The prediction error using a map h for $x \in X$ is then given as

$$E_h(x) \coloneqq L(h(x), \underline{f}(x)) \,, \tag{2.5}$$

which is a random variable for fixed $h \in \mathcal{H}$.

Accordingly, we obtain the *generalization error* $E^\star(h)$ by

$$E^\star(h) \coloneqq \mathbb{E}[E_h] \,. \tag{2.6}$$

As the distribution P on X is usually unknown, the generalization error cannot be computed in practice. Therefore, an empirical error is used as a proxy on a labeled sample set $S \subset X$. The *empirical error* for $h \in \mathcal{H}$ is then given as

$$\widehat{E}_S(h) \coloneqq \frac{1}{|S|} \sum_{x \in S} L(h(x), \underline{f}(x)) \,. \tag{2.7}$$

For fixed $h \in \mathcal{H}$, the expected value of the empirical error equals the generalization error if samples are drawn independently and identically distributed, which motivates optimizing the empirical error.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

To see this, we fix a map $h \in \mathcal{H}$ and a number of samples r. Then, $\widehat{E}_S(h)$ is a random variable defined on X^r: Let $E_h^{(i)} := E_h \circ \pi_i : X^r \to \mathbb{R}$, where π_i is the projection on the ith component. Then, for $S \in X^r$ drawn independently and identically distributed, we have

$$\widehat{E}_S(h) = \frac{1}{r} \sum_{i=1}^{r} E_h^{(i)}(S)\,, \tag{2.8}$$

considered as a random variable. By assumption on S, we conclude

$$\begin{aligned} \mathbb{E}\Big[\widehat{E}_S(h)\Big] &= \mathbb{E}\left[\frac{1}{r}\sum_{i=1}^{r} E_h^{(i)}\right] = \frac{1}{r}\sum_{i=1}^{r} \mathbb{E}\Big[E_h^{(i)}\Big] \\ &= \mathbb{E}\Big[E_h^{(1)}\Big] = \mathbb{E}[E_h] \\ &= E^\star(h)\,, \end{aligned}$$

as the expected value is linear and $\mathbb{E}[E_h] = \mathbb{E}\Big[E_h^{(i)}\Big]\ \forall i \in [n]$.

A commonly used approach in supervised learning, which we follow in this work, is trying to minimize the empirical error

$$h_S^\star = \operatorname*{arg\,min}_{h \in \mathcal{H}} \widehat{E}_S(h)\,, \tag{2.9}$$

which is called *empirical risk minimization*. Several bounds are known [128] that relate $E^\star(h_S^\star)$ with the minimal generalization error $\inf_{h \in \mathcal{H}} E^\star(h)$, thus justifying the approach.

When the empirical risk is small but the generalization error is relatively big, it means that a machine learning algorithm is not able to generalize, which is also called *overfitting*. This happens if the hypothesis set contains overly complex maps, *i.e.*, the degrees of freedom of a map $h \in \mathcal{H}$ is too high, compared to the number of training data.

Supervised learning also enables modelling that there is uncertainty in the labels. In this case, a probability distribution over $X \times Y$ is considered. In general, supervised learning can also be understood as maximizing the conditional distribution $\mathsf{P}(y \mid x)$, given $x \in X$.

2.4.2 Logistic regression

A supervised learning model that is frequently used in this work is *logistic regression* [129]. Consider a probability distribution over $X \times Y$ with $X \subseteq \mathbb{R}^n$ and binary label space $Y = \{0, 1\}$. For $\mathbf{x} \in X$, we define $p_{\mathbf{x}} = \mathsf{P}(Y = 1 \mid X = \mathbf{x})$. Then, $1 - p_{\mathbf{x}} = \mathsf{P}(Y = 0 \mid X = \mathbf{x})$. Assuming $p_{\mathbf{x}} \in (0, 1)$, the *odd* for an event is given by

$$o(p_{\mathbf{x}}) := \frac{p_{\mathbf{x}}}{1 - p_{\mathbf{x}}} \in (0, \infty)\,. \tag{2.10}$$

The concept behind logistic regression is to describe probabilities in terms of linear regression. The odd must thus be converted to be applicable. This is done using the *log-odd* $\log(o(p_{\mathbf{x}})) \in (-\infty, \infty)$, as it allows computing weights $\mathbf{w} \in \mathbb{R}^n$ and bias $b \in \mathbb{R}$ such that

$$\log(o(p_{\mathbf{x}})) = \langle \mathbf{w}, \mathbf{x} \rangle + b\,. \tag{2.11}$$

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

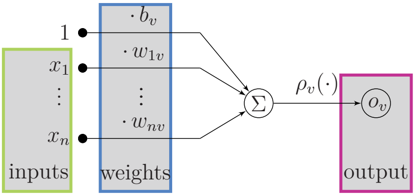


Figure 2.5: Visualization of a neuron o_v for an input vector $(x_1, \dots, x_n)$.

The probability of $\mathbf{x} \in X$ having label 1 is then

$$p_{\mathbf{x}} = \frac{1}{1 + \exp(-\langle \mathbf{w}, \mathbf{x} \rangle - b)} . \tag{2.12}$$

The best-fitting parameters $\mathbf{w}$ and b can be obtained from a maximum likelihood estimation applied to a training set. Usually, this is done via the log-likelihood which can be computed using gradient descent or iteratively reweighted least squares. This thesis uses the software implementation LIBLINEAR [74] in Chapter 3 to learn the probabilities of two detections belonging to the same person.

2.4.3 Neural networks

Machine learning and computer vision have seen significant improvements in recent years due to the latest advances in neural networks [130–133]. In this work, neural networks are employed for supervised learning; to perform object detection, person re-identification, and feature computation (see Chapter 3 and Chapter 4). We briefly introduce the basic concepts of neural networks and refer the interested reader to Goodfellow *et al.* [134] for a more comprehensive introduction to the topic.

Definition 2.17. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an acyclic directed graph with edge weights $\mathbf{w} = (w_{uv})_{uv \in \mathcal{E}}$ and vertex weights $\mathbf{b} = (b_v)_{v \in \mathcal{V}}$. For $v \in \mathcal{V}$, let $\rho_v : \mathbb{R} \to \mathbb{R}$ be a function. For $\mathbf{x} \in \mathbb{R}^n$, we recursively define a mapping $o_v : \mathbb{R}^n \to \mathbb{R}$ via

$$o_v(\mathbf{x}) := \begin{cases} \rho_v \left(b_v + \sum_{u \in N^-_{\mathcal{G}}(v)} w_{uv} o_u(\mathbf{x}) \right) & \text{if } d^-_{\mathcal{G}}(v) > 0 \,, \\ x_v & \text{otherwise.} \end{cases} \tag{2.13}$$

The map o_v is called (artificial) *neuron*, b_v is a *bias* and ρ_v an *activation function*. Finally, $(\mathcal{V}, \mathcal{E}, \{o_v \mid v \in \mathcal{V}\})$ is called *feedforward neural network*.

Each neuron o_v aggregates $d^-_{\mathcal{G}}(v)$ many weighted input signals. Before passing the values to the activation function ρ_v, a bias b_v is added. This computational process of a neuron is illustrated in Figure 2.5.

A feedforward neural network consists of multiple interconnected neurons. In the following, we shall denote such a network simply as a neural network.

Neural networks are divided into *layers*, where each layer consists of neurons with similar structure: The neurons have a decomposition $\mathcal{V} = \mathcal{V}_1 \sqcup \ldots \sqcup \mathcal{V}_{n_L}$ with n_L being the number

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

of layers such that $\forall l \in [2:n_L]\, \forall v, u \in \mathcal{V}_l : \rho_v \equiv \rho_u$, *i.e.*, neurons of the same layer have the same activation function. Furthermore, $\forall l \in [2:n_L]\, \forall v \in \mathcal{V}_l : N^-_{\mathcal{G}}(v) \subseteq \mathcal{V}_1 \cup \ldots \cup \mathcal{V}_{l-1}$, *i.e.*, each node has incoming edges only from previous layers. Usually, regularity is assumed per layer, *i.e.*, $\forall l \in [n_L] \forall v, u \in \mathcal{V}_l : d^-_{\mathcal{G}}(v) = d^-_{\mathcal{G}}(u)$. We denote the number of neurons in layer $l \in [n_L]$ as $n_l := |\mathcal{V}_l|$. Given a neural network with n_L layers, using the neural network of all layers up to layer $l \leq n_L$ is called a *backbone*. In practice, neural networks consisting of more than 150 layers [132, 135] have been successfully used for computer vision tasks. Observations have shown [136] that earlier layers are typically sensitive to low-level image features while later layers compose these features to form more complex image features.

In the subsequent part of this section, it is assumed that edges exist only between consecutive layers. Then, for fixed $\mathbf{x} \in \mathbb{R}^{n_1}$, the output $o_v(\mathbf{x})$ at layer $l \in [2:n_L]$ depends entirely on the vector $\mathbf{o}^{(l-1)} := \mathbf{o}^{(l-1)}(\mathbf{x}) := (o_v(\mathbf{x}))_{v \in \mathcal{V}_{l-1}} \in \mathbb{R}^{n_{l-1}}$ of layer $l-1$. As a result, such neural networks are compositions of layer-wise functions.

2.4.3.1 Layers

This work focuses on a specific class of feedforward neural networks called *Convolutional Neural Network*s (CNNs) [137, 138], which have turned out to be very effective, improving the accuracy of many machine learning tasks considerably, especially in the field of image analysis. In this work, CNNs are used to obtain accurate people detections as a basis for the proposed tracking methods. Also, CNNs are employed to regress the orientation of a person in Section 3.5. Finally, CNNs provide very reliable pairwise features between detections (see Chapter 4).

CNNs use certain layers, termed *fully connected*, *convolution*, and *pooling* layers. We present these layers subsequently, as they are used throughout this work. In the following, let a layer $l \in [2:n_L]$ be fixed.

Fully-connected layer. Layer l is called *fully-connected* if each node of layer l is connected to all nodes of the previous layer $l-1$ so that $N^-_{\mathcal{G}}(v) = \mathcal{V}_{l-1}$ for all $v \in \mathcal{V}_l$. Thus for each node $v \in \mathcal{V}_l$, bias vector $\mathbf{b} \in \mathbb{R}^{n_l}$, weight matrix $\mathbf{W} \in \mathbb{R}^{n_l \times n_{l-1}}$, and activation function ρ, we have

$$\mathbf{o}^{(l)} = \boldsymbol{\rho}(\mathbf{W}\mathbf{o}^{(l-1)} + \mathbf{b}) \tag{2.14}$$

using the mapping $\boldsymbol{\rho} : \mathbb{R}^{n_l} \to \mathbb{R}^{n_l}, \mathbf{x} \mapsto (\rho(x_1), \ldots, \rho(x_{n_l}))$.

In theory, feedforward networks using only fully connected layers can represent most functions used in computer vision, which is known as the universal approximation theorem:

Theorem 2.18. *Let $\mathcal{H}^{n_3}_{n_1}$ be the set of all feedforward neural networks consisting of 3 layers and let each layer $l \in [2:3]$ be fully connected. Let $\rho^{(2)}$ be a continuous activation function used for all neurons of layer 2, and let the identity be the activation function used for all neurons of layer 3. Let $K \subseteq \mathbb{R}^{n_1}$ be compact. Then, $\mathcal{H}^{n_3}_{n_1}$ is dense in $\{f : K \to \mathbb{R}^{n_3} \mid f \text{ continuous}\}$ with respect to the supremum norm if and only if $\rho^{(2)}$ is non-polynomial.*

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Proof. See Pinkus [139]. □

Several other variants of the universal approximation theorem have been established (*e.g.*, with an arbitrary number of layers [140]), showing that feedforward neural networks have excellent theoretical approximation properties. Note however that (i) the proof of Theorem 2.18 is non-constructive and (ii) the approximation property is guaranteed only within a bounded set. Yet, the universal approximation theorem shows that in theory, there is great flexibility in selecting an activation function. The *sigmoid function* $\sigma(x) \coloneqq \frac{1}{1+\exp(-x)}$, the hyperbolic tangent $\tanh(x)$, and the so-called *Rectified Linear Unit* (ReLU) defined by $relu(x) \coloneqq \max\{0, x\}$ are common choices. Note that using a single neuron with the sigmoid activation function is nothing else but logistic regression.

A drawback of using fully connected layers is that the number of weights that need to be determined and the number of computations that need to be performed can become huge, resulting in overfitting as well as computational issues.

2D convolution layer. To reduce computational costs and avoid overfitting issues, convolutional layers are used. Assuming that the nodes of layer $l-1$ are arranged in a 2D grid structure, layer $l-1$ is viewed as an image $\mathbf{I} \in \mathbb{R}^{w \times h \times n_c}$ of dimension $w \times h$ with n_c channels. A 2D convolution layer then applies to each channel of $\mathbf{I}$ a kernel via 2D convolution and sums up the results across the channels. Simultaneously performing n_{out} 2D convolutions results in an output that can be viewed as an image with n_{out} channels. For $c_{in} \in [n_c]$, we reshape $\mathbf{I}_{[:,:,c_{in}]}$ as a vector $\mathbf{o}_{c_{in}}^{(l-1)} \in \mathbb{R}^{w \cdot h}$. The output is then given by

$$\mathbf{o}^{(l)} = \left(\boldsymbol{\rho}\Big(\sum_{c_{in}=1}^{n_c} \mathbf{W}_{c_{in},c_{out}} \mathbf{o}_{c_{in}}^{(l-1)} + b_{c_{out}} \mathbf{1} \Big) \right)_{c_{out} \in [n_{out}]} . \tag{2.15}$$

A Toeplitz matrix $\mathbf{W}_{c_{in},c_{out}} \in \mathbb{R}^{w'h' \times wh}$ is used to perform the convolution. The bias $b_{c_{out}} \mathbf{1} \in \mathbb{R}^{w'h'}$ with $b_{c_{out}} \in \mathbb{R}$ is constant for each output channel.

By definition of a Toeplitz matrix (or equivalently a convolution), multiple edge weights of the neural network are constrained to share the same value, which is called *weight-sharing*. Consequently, the degrees of freedom of a neural network are heavily reduced compared to using fully connected layers, making the network less prone to overfitting issues. At the same time, convolutions are established methods to retrieve image features in computer vision [141], mainly due to their translation equivariance properties. Hence, learning kernel weights has become essential for the successful application of neural networks in computer vision tasks [132, 133].

2D pooling layer. Assuming that layer $l-1$ is arranged in a 2D grid structure, it can be viewed as an image $\mathbf{I} \in \mathbb{R}^{w \times h \times n_c}$ of dimension $w \times h$ with n_c channels. Pooling layers reduce the spatial dimension of $\mathbf{I}$ so that the resulting output can be viewed as an image $\mathbf{I}' \in \mathbb{R}^{w' \times h' \times n_c}$ of spatial dimension $w' \times h' < w \times h$.

A neighborhood (*e.g.*, 3×3) is fixed, which we denote as N. For an image position $\mathbf{p} \in [w] \times [h]$ and $N(\mathbf{p}) \coloneqq \mathbf{p} + N$, a 2D pooling layer outputs for each channel $c \in [n_c]$ the value $o_{\mathbf{p},c} = f\left(\mathbf{I}(N(\mathbf{p}), c)\right)$, where f either computes the mean value or max value of $\mathbf{I}(N(\mathbf{p}), c)$, which is called *average pooling* and *max pooling*, respectively. To reduce

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

the spatial dimension, pooling is computed on a proper subset $P \subsetneq [w] \times [h]$ of image positions.

As the information within a neighborhood is aggregated in a pooling layer, a neural network becomes robust to certain translations. Pooling can also be understood as a form of noise suppression, *e.g.*, when the important information is contained in a local maximum. Most importantly, a pooling layer reduces computational costs and mitigates overfitting.

2.4.3.2 Training of Neural Networks

In this work, neural networks are applied using the supervised learning setting (see Section 2.4.1). Accordingly, given a training set S, the empirical error $\widehat{E}_S$, according to Eq. (2.7), needs to be minimized. The hypothesis set $\mathcal{H}(\mathcal{G})$ is specified by fixing a neural network graph $\mathcal{G}$ with corresponding activation functions. Each $h \in \mathcal{H}(\mathcal{G})$ corresponds to a specific assignment of weights $\mathbf{w}$ and biases $\mathbf{b}$ to $\mathcal{G}$ so that we may denote a hypothesis, *i.e.*, a neural network, as $h_{\mathbf{w},\mathbf{b}} \in \mathcal{H}(\mathcal{G})$. In order to train a neural network, a loss L needs to be defined so that the empirical error can be minimized.

Loss. In the following, we consider a mapping $\underline{f} : X \to Y$ that we want to learn using supervised learning. Let $S \subset X$ be a training set and let $h_{\mathbf{w},\mathbf{b}} \in \mathcal{H}(\mathcal{G})$ be a neural network.

For binary classification, *i.e.*, $Y = \{0, 1\}$, the *Binary Cross-Entropy* (BCE) loss L_{BCE} is often used, which for $\mathbf{x} \in X, y \in Y$ is given as

$$L_{\text{BCE}}(h_{\mathbf{w},\mathbf{b}}(\mathbf{x}), y) = -y \log(h_{\mathbf{w},\mathbf{b}}(\mathbf{x})) + (1 - y) \log(1 - h_{\mathbf{w},\mathbf{b}}(\mathbf{x}))\,. \tag{2.16}$$

It is assumed that $h_{\mathbf{w},\mathbf{b}}(\mathbf{x}) \in (0, 1)$ for all $\mathbf{x} \in X$. The binary cross-entropy loss is a measure of dissimilarity between the distribution of the true labels and the distribution given by the neural network $h_{\mathbf{w},\mathbf{b}}$. It is used in Section 4.7.3 to train a neural network that classifies whether a pair of detections belongs to the same person.

In the case of multi-class classification with n_C classes, *i.e.*, $Y = [n_C]$, a neural network $h \in \mathcal{H}$ maps an input $\mathbf{x}$ to $h_{\mathbf{w},\mathbf{b}}(\mathbf{x}) = (h^{(1)}_{\mathbf{w},\mathbf{b}}(\mathbf{x}), \ldots, h^{(n_C)}_{\mathbf{w},\mathbf{b}}(\mathbf{x})) \in (0, 1)^{n_C}$. As loss, the *Cross Entropy* (CE) can be used, which is given via

$$L_{\text{CE}}(h_{\mathbf{w},\mathbf{b}}(\mathbf{x}), y) = -\sum_{c=1}^{n_C} [y = c] \log(h^{(c)}_{\mathbf{w},\mathbf{b}}(\mathbf{x}))\,, \tag{2.17}$$

where $[y = c]$ denotes the Iverson bracket. It is applied in Section 2.7.2.2 to train robust appearance features used to re-identify persons.

For regression problems, we consider $Y \subseteq \mathbb{R}^n$. In this case, minimizing the *Mean Squared Error* (MSE) loss L_{MSE} given by

$$L_{\text{MSE}}(h_{\mathbf{w},\mathbf{b}}(\mathbf{x}), \mathbf{y}) = \|h_{\mathbf{w},\mathbf{b}}(\mathbf{x}) - \mathbf{y}\|^2 \tag{2.18}$$

results in minimizing the average distance between the predicted and ground truth vector. Neural network based object detectors (see Section 2.7.1) are trained to

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

regress coordinates of bounding boxes using the L_{MSE} loss. Alternatively, the *Cosine Similarity* (CS) can be maximized, which is given by

$$L_{\text{CS}}(h_{\mathbf{w},\mathbf{b}}(\mathbf{x}),\mathbf{y}) = \frac{\langle h_{\mathbf{w},\mathbf{b}}(\mathbf{x}),\mathbf{y}\rangle}{\|h_{\mathbf{w},\mathbf{b}}(\mathbf{x})\|\,\|\mathbf{y}\|}. \tag{2.19}$$

It then minimizes the average angle between predicted and ground truth vectors, which is used in Section 3.5.2.2 to regress the orientation of a person.

Computing optimal weights. Finding the optimal solution $h^\star \in \mathcal{H}(\mathcal{G})$ typically poses a non-linear and non-convex optimization problem (depending on the layers and activation functions). Thus, we have to minimize the empirical error $\hat{E}_S$ which we consider as a map $\hat{E}_S : \mathbb{R}^n \to \mathbb{R}$, $\hat{E}_S(\mathbf{w},\mathbf{b}) := \hat{E}_S(h_{\mathbf{w},\mathbf{b}})$, where we assume that each $h_{\mathbf{w},\mathbf{b}} \in \mathcal{H}(\mathcal{G})$ is parameterized by a vector $(\mathbf{w},\mathbf{b}) \in \mathbb{R}^n$ and $\hat{E}_S$ is assumed to be continuously differentiable.

Variants of the gradient descent algorithm are usually applied to minimize the function. Subsequently, we briefly introduce the main concepts of gradient descent.

Definition 2.19. Let $\mathbf{d} \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$. Then $\mathbf{d}$ is a *descent direction* at $\mathbf{x} \in \mathbb{R}^n$ if there exists a value $\Lambda \in \mathbb{R}_{>0}$ such that for all $\lambda \in (0,\Lambda)$, $f(\mathbf{x}+\lambda\mathbf{d}) < f(\mathbf{x})$.

The value λ is called *step-size* or in the context of machine learning *learning rate*. The following theorem gives a sufficient condition for a descent direction.

Theorem 2.20. *If* $f : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable and* $\nabla f(\mathbf{x})\mathbf{d} < 0$*, then* $\mathbf{d}$ *is a descent direction for* $\mathbf{x} \in \mathbb{R}^n$*.*

Proof. For $F(\lambda) := f(\mathbf{x}+\lambda\mathbf{d})$, we have $\frac{dF}{d\lambda}(\lambda) = \nabla f(\mathbf{x}+\lambda\mathbf{d})\mathbf{d}$. The statement then follows from $0 > \nabla f(\mathbf{x})\mathbf{d} = \frac{dF}{d\lambda}(0) = \lim_{\substack{u\to 0\\ u>0}} \frac{F(u)-F(0)}{u}$. ☐

Thus, if $\hat{\mathbf{w}} := (\mathbf{w},\mathbf{b})$ is not a *stationary point*, *i.e.*, $\nabla\hat{E}_S(\hat{\mathbf{w}}) \neq \mathbf{0}$, then $\mathbf{d} := -(\nabla\hat{E}_S(\hat{\mathbf{w}}))^\intercal$ satisfies the conditions of Theorem 2.20, since

$$\nabla\hat{E}_S(\hat{\mathbf{w}})\mathbf{d} = -||\nabla\hat{E}_S(\hat{\mathbf{w}})||^2 < 0\,. \tag{2.20}$$

To find optimal weights and biases for $\hat{E}_S$, a current solution $\hat{\mathbf{w}}$ is thus moved towards the negative gradient of $\hat{E}_S$ at $\hat{\mathbf{w}}$.

In order to apply gradient descent, the neural network and the loss function must be differentiable with respect to the weights and biases. Different algorithms have been developed to find the corresponding learning rate, like the Armijo-rule that performs an additional 1D line search. It can be shown that gradient descent with the Armijo-rule for a continuously differentiable function either terminates on a stationary point or creates a strictly monotonic decreasing sequence such that each accumulation point is a stationary point. More details about gradient descent can be found in any textbook about non-linear optimization, *e.g.*, Ulbrich *et al.* [142].

However, computing the gradient $\nabla\widehat{E}_S$ becomes very expensive if the training set S is large. It is therefore common to apply *stochastic gradient descent* [143]. The training

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

set S is randomly partitioned into n sets $B_1, \dots, B_n$ of equal size, called *mini-batches*. For each mini-batch, the parameters are updated according to the gradient. Stochastic gradient descent thus employs an approximation of the true gradient $\nabla \widehat{E}_S$, using the gradients $\nabla \widehat{E}_{B_k}$. The entire algorithm is sketched in Algorithm 2.2.

Algorithm 2.2: Stochastic gradient descent

```
Input   : Training set $S$.
Output  : Optimized weights $(\mathbf{w}, \mathbf{b})$.

Initialize weights $(\mathbf{w}, \mathbf{b})$
while stopping criteria not true do
    Set learning rate $\lambda$
    Randomly partition $S$ into mini-batches $B_1, \dots, B_n$ of equal size
    for $k \in [n]$ do
        $(\mathbf{w}, \mathbf{b}) \leftarrow (\mathbf{w}, \mathbf{b}) - \lambda \nabla \widehat{E}_{B_k}(\mathbf{w}, \mathbf{b})$
    end for
end while
```

Note that the gradients are typically computed efficiently and automatically using the automatic differentiation method [144] in reverse mode, which is also called *backpropagation* [145] in the context of neural networks.

For stochastic gradient descent, it has been shown that convergence is almost sure guaranteed to a critical point by choosing a fast-enough diminishing step-size [146, 147]. In addition, saddle points are avoided almost sure even if the objective function is non-convex (but satisfies certain conditions) [148].

2.4.3.3 Regularization

A neural network can contain a huge[1] number of variables that need to be determined by the optimization algorithm. Therefore, neural networks are prone to overfitting. Regularization techniques help to avoid these problems.

Batch normalization. A frequently used technique is to apply *batch normalization* [150] on the output of a layer, which is used in the tracking method of Chapter 4. Batch normalization enables faster convergence during training and helps mitigate overfitting, thus having a regularizing effect that leads to higher accuracies. It performs per coordinate a z-score normalization for each mini-batch during training. During inference, population statistics are used for the normalization.

In more detail for a mini-batch $B \subset \mathbb{R}^d$, the empirical mean $\mu(B)$ and variance $\sigma^2(B)$ of B is given as

$$\boldsymbol{\mu}(B) = \frac{1}{|B|} \sum_{\mathbf{x} \in B} \mathbf{x}, \quad \text{and } \boldsymbol{\sigma}^2(B) = \left(\frac{1}{|B|} \sum_{\mathbf{x} \in B} (x_j - \mu(B)_j)^2 \right)_{j \in [d]}. \tag{2.21}$$

---

[1] As an example, the GPT-3 network used in natural language processing has $\approx 175 \cdot 10^9$ parameters [149].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

For $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d$ and $\mathbf{x} \in B$, normalization is performed per coordinate $j \in [d]$, *i.e.*,

$$\hat{x}_j := \alpha_j \frac{x_j - \mu(B)_j}{\sqrt{\sigma^2(B)_j + \epsilon}} + \beta_j \,, \tag{2.22}$$

where $\epsilon > 0$ is a small number used to avoid numerical issues. The vectors $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d$ are viewed as variables that are optimized during training. The vector $\hat{\mathbf{x}} = (\hat{x}_j)_{j \in [d]}$ is passed to the next layer.

During inference given $\mathbf{x} \in \mathbb{R}^d$, the transformation is computed as

$$\hat{x}_j = \alpha_j \frac{x_j - \mu_j}{\sqrt{\sigma_j^2 + \epsilon}} + \beta_j \,, \tag{2.23}$$

where μ_j and σ_j^2 are determined by the empirical mean $\boldsymbol{\mu}(B)$ and variance $\boldsymbol{\sigma}^2(B)$ of all batches B via moving average during training.

Dropout. Another regularization technique called *Dropout* [151] introduces artificial noise during training. Applied to the output of layer l, the output of each neuron is set to 0 with probability $p \in (0, 1)$. Dropout effectively mitigates overfitting as the neural network is trained to output the correct results based on a sparse signal flow and to compensate for errors in the input. This technique is used in Section 3.5 to avoid overfitting while regressing the orientation of a person.

2.5 Computational Complexity Theory

While HO-MOT is conceptually able to deliver very accurate results, solving the underlying optimization poses a challenging problem that may prevent an HO-MOT approach from being applicable in practice. To characterize the difficulty of such optimization problems, this section briefly introduces computational complexity theory, following the notation and statements presented by Bovet *et al.* [152].

Computational complexity theory is applicable to problems that are formalized in terms of a formal language.

Definition 2.21. A finite set $\Sigma \neq \emptyset$ is called *alphabet* and $L \subseteq \Sigma^*$ a *formal language*. Thereby, $\Sigma^* := \bigcup_{k=0}^{\infty} \Sigma^k$ and $\Sigma^0 := \{\epsilon\}$ with ϵ being the *empty word*. An element $x = (x_1, \ldots, x_k) \in \Sigma^*$ is a *word* of *length* k. We shall write $x = x_1 \ldots x_k$ and $|x| := k$. Note that $|\epsilon| = 0$.

Words of a language encode problem instances and computational complexity theory categorizes languages into difficulty classes. Of special interest are languages that describe decision problems.

Definition 2.22. A *decision problem* is a tuple $\Pi = (I, S, \pi)$, where $I \subseteq \Sigma^*$ is a set of words representing problem instances. The map S assigns a word $i \in I$ to a finite set $S(i) \subseteq \Sigma^*$ representing candidate solutions. Finally, $\pi : I \times S \to \{\text{true}, \text{false}\}$ is a map so that $\pi(i, y)$ is true if and only if $y \in S(i)$ is a valid solution for $i \in I$. Further definitions for $i \in I$ are as follows:

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

- The *valid set* is given by $V(i) \coloneqq \{y \mid y \in S(i) \wedge \pi(i, y)\}$.
- *Solving* the decision problem is to decide if $V(i)$ is non-empty.

Finally, the language of Π is defined as $L_\Pi \coloneqq \{j \in I \mid V(j) \neq \emptyset\}$.

A decision problem can be understood as answering if at least one valid solution to a given problem instance exists.

Example 2.23. Given a complete bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, w)$, the question of whether an optimal matching $\mathcal{E}' \subseteq \mathcal{E}$ exists with matching costs $w(\mathcal{E}') \geq k$ poses a decision problem: A word of the set I describes (in some encoding) a weighted complete bipartite graph and k. For $i \in I$, the set $S(i)$ encodes all subgraphs $(\mathcal{V}', \mathcal{E}') \subseteq \mathcal{G}$ and $\pi(i, y)$ is true if and only if y encodes a subgraph $(\mathcal{V}', \mathcal{E}') \subseteq \mathcal{G}$ with $\mathcal{V} = \mathcal{V}'$ and $\mathcal{E}'$ is an optimal matching with $w(\mathcal{E}') \geq k$.

In the following, we implicitly assume any decision problem Π to be transformed to the language L_Π (see Definition 2.22).

Fundamental for assessing a language is the concept of a Turing machine, which is a mathematical model of a computing machine.

Definition 2.24. A (deterministic) *k-tape Turing machine* (with $k \in \mathbb{N}$) consists of

- a set of states Q,
- an alphabet Σ including a special symbol $\square$ called *blank*,
- a finite set I of tuples $(q_1, \mathbf{s}_1, \mathbf{s}_2, \mathbf{m}, q_2)$, where $q_1, q_2 \in Q$, $\mathbf{s}_1 \in \Sigma^k$, $\mathbf{s}_2 \in (\Sigma \setminus \{\square\})^k$, $\mathbf{m} \in \{L, R, S\}^k$,
- an *initial state* $\underline{q} \in Q$,
- a set of *final states* $F \subseteq Q$.

For all $\mathbf{i} = (q_1^{(i)}, \mathbf{s}_1^{(i)}, \mathbf{s}_2^{(i)}, \mathbf{m}^{(i)}, q_2^{(i)}), \mathbf{j} = (q_1^{(j)}, \mathbf{s}_1^{(j)}, \mathbf{s}_2^{(j)}, \mathbf{m}^{(j)}, q_2^{(j)}) \in I$ it must hold that if $q_1^{(i)} = q_1^{(j)}$ and $\mathbf{s}_1^{(i)} = \mathbf{s}_1^{(j)}$, then $\mathbf{i} = \mathbf{j}$.

A k-tape Turing machine contains k *memory tapes*. A *tape* is a tuple (τ, h), where $\tau : \mathbb{N} \to \Sigma$ stores a character at each position (thus representing memory) and $h \in \mathbb{N}$ is the current read/write position (called *head position*). L, R, and S shift the head position h by one to the left, by one to the right, or keep the current position, respectively. All tapes are initialized with the blank character and head position $h = 1$.

A Turing machine T is initialized with $\underline{q}$ as state and initialized tapes. To compute $T(x)$, the word x is written on one tape, starting at write position 1.

Definition 2.25. One *step* of a Turing machine reads the characters of the tapes at the respective head positions $\mathbf{h} \in \mathbb{N}^k$, resulting in $\mathbf{s} \in \Sigma^k$. Given the current state $q \in Q$, a tuple $(q, \mathbf{s}, \mathbf{s}', \mathbf{m}, q') \in I$ is applied, which means that the tape information $\mathbf{s}$ is replaced by $\mathbf{s}'$ at $\mathbf{h}$, the head positions are updated according to $\mathbf{m}$, and q' is set as the new state.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Thus, a Turing machine performs for a given input $x \in \Sigma^*$ a sequence of (possibly infinitely many) steps. If T reaches a final state after finitely many steps, it *halts*.

Definition 2.26. A Turing machine T is an *acceptor* if, for all $x \in \Sigma^*$, whenever $T(x)$ results in a finite sequence of steps, the last state is a final state, which must be either an *accept* or *reject* state. An acceptor Turing machine T *accepts* a language L if the set of all inputs, where T halts in an accept state, equals L.

Definition 2.27. A Turing machine T is a *transducer* if, for all $x \in \Sigma^*$, whenever $T(x)$ results in a finite sequence of steps, the last state is a final state, called the *terminate* state. In addition, it must return an output $y \in \Sigma^*$ written on one of its tapes.

Most importantly, an algorithm expressed in pseudo pascal code can be performed by a Turing machine and vice versa. Essentially, computational complexity theory is about categorizing how efficiently a Turing machine can accept a language.

For the decision problem of Example 2.23, a Turing machine that accepts the corresponding language outputs for each complete bipartite graph $\mathcal{G}$ that has an optimal matching with costs $c \geq k$, the accept state.

Using a language and a Turing machine, we can define computable functions.

Definition 2.28. A function f defined on $L \subseteq \Sigma^*$ is *computable* if a transducer Turing machine T exists such that, for all $x \in L$, $T(x)$ reaches the terminate state and outputs $f(x)$.

In order to categorize decision problems according to their complexity, the following definition is crucial:

Definition 2.29. A language $L \subseteq \Sigma^*$ has deterministic polynomial time complexity, denoted by $\mathcal{P}$, if there exists a Turing machine T that (i) accepts L, (ii) halts for every input $x \in \Sigma^*$, and (iii) for all $x \in \Sigma^*$, the number of steps required[2] to halt is polynomial in the length $|x|$.

Decision problems belonging to the complexity class $\mathcal{P}$ are considered to be efficiently solvable. An example of a decision problem belonging to $\mathcal{P}$ is Example 2.23, according to Theorem 2.15. Efficiently computable functions are defined in terms of a transducer Turing machine, accordingly.

Definition 2.30. The class $\mathcal{FP}$ is the class of all functions that are computable in polynomial runtime by a deterministic transducer Turing machine.

This allows comparing the complexity of two languages $L_1, L_2 \subseteq \Sigma^*$ (and in particular languages of decision problems).

Definition 2.31. Given two languages L_1 and L_2, the language L_1 is *polynomially reducible* to L_2, denoted as $L_1 \leq_r L_2$, if a function $f \in \mathcal{FP}$ exists such that $x \in L_1 \iff f(x) \in L_2$.

---

[2] To be precise, the step counting function must be computable by a Turing machine.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Assume that $L_1 \leq_r L_2$ via a function f and let $x \in \Sigma^*$ be an input word. If $L_2 \in \mathcal{P}$, then a Turing machine exists that terminates for each input $f(x)$ in polynomial runtime (with respect to $f(x)$) and accepts L_2. As $f(x) \in L_2$ can be decided in polynomial runtime, it allows deciding in polynomial runtime if $x \in L_1$, since $L_1 \leq_r L_2$.

Thus given two decision problems L_1, L_2 with $L_1 \leq_r L_2$, it follows that if L_2 can be solved efficiently, then the decision problem L_1 can be solved efficiently too. Thus, if $L_1 \leq_r L_2$, then the decision problem L_2 is at least as difficult as L_1.

Besides the complexity class $\mathcal{P}$, which categorizes languages in terms of their efficiency in accepting a language, another characterization is how efficiently a solution can be verified.

Definition 2.32. A language $L \subseteq \Sigma^*$ is in the *non-deterministic polynomial time* complexity class $\mathcal{NP}$ if it is verifiable in polynomial time, *i.e.*, if a language $L_{\text{check}} \in \mathcal{P}$ exists and a polynomial p such that $L = \{x \mid \exists y : xy \in L_{\text{check}} \land |y| \leq p(|x|)\}$.

The interpretation is as follows: A word x encodes a problem instance and the word y a possible solution. Since $L_{\text{check}} \in \mathcal{P}$, it is decidable in polynomial time if the solution is valid, and the length of the solution word y depends polynomially on (and is upper bounded by) the length of the word x.

The next example shows a decision problem that can easily be verified to be in $\mathcal{NP}$.

Example 2.33 (3-SAT)**.** Given a binary variable x, the expressions x and its negation $\overline{x}$ are the *literals* that can be created from x. A *clause* is a formula $f = f_1 \vee \ldots \vee f_n$, where each f_i is a literal. Let $G = g_1 \wedge \ldots \wedge g_m$ be a formula, where each g_j is a clause of at most 3 literals. The *3-SAT* decision problem is to decide whether there exists an assignment of the involved binary variables such that G holds.

Until today, no algorithm with polynomial runtime in the input length has been found to solve the 3-SAT problem for arbitrary input.

Remark 2.34. A solution that can be constructed in polynomial time can also be verified in polynomial time so that $\mathcal{P} \subseteq \mathcal{NP}$ holds. Yet until today, whether $\mathcal{P} = \mathcal{NP}$ or $\mathcal{P} \neq \mathcal{NP}$ holds is an open question[3].

Using the formalization of reduction, we can characterize languages as being at least as difficult as any language of $\mathcal{NP}$.

Definition 2.35. A language L is $\mathcal{NP}$*-hard* if for any language $L' \in \mathcal{NP}$, the reduction $L' \leq_r L$ holds.

The most difficult languages of $\mathcal{NP}$ are thus $\mathcal{NP}$-hard.

Definition 2.36. A language L is $\mathcal{NP}$*-complete* if $L \in \mathcal{NP}$ and L is $\mathcal{NP}$-hard.

An example of an $\mathcal{NP}$-complete decision problem is the 3-SAT problem defined in Example 2.33 (see Karp [154]).

---

[3] The so-called $\mathcal{P}$ *vs.* $\mathcal{NP}$ problem is a millennium prize problem [153].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

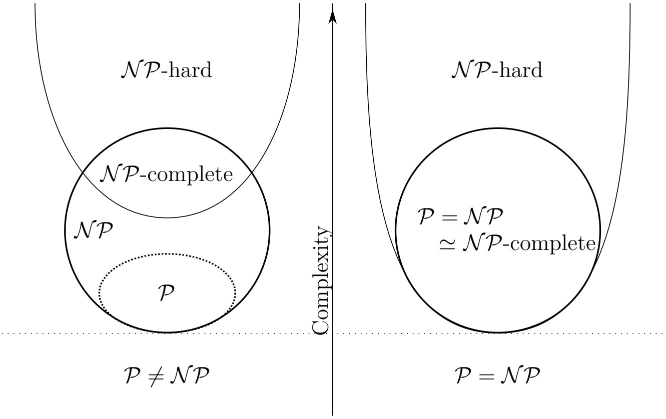

$\mathcal{NP}$-hard
$\mathcal{NP}$-complete
$\mathcal{NP}$
$\mathcal{P}$
Complexity
$\mathcal{NP}$-hard
$\mathcal{P} = \mathcal{NP}$
$\simeq \mathcal{NP}$-complete
$\mathcal{P} \neq \mathcal{NP}$
$\mathcal{P} = \mathcal{NP}$

Figure 2.6: Relation of complexity classes in case that $\mathcal{P} \neq \mathcal{NP}$ holds (left) and in the case of equality (right) [159][4]. In the left case, the classes $\mathcal{P}$ and $\mathcal{NP}$-complete are disjoint. In the case of equality, the classes $\mathcal{P}$, $\mathcal{NP}$, and $\mathcal{NP}$-complete are not distinguishable.

Thus, if an efficient algorithm for an $\mathcal{NP}$-hard problem was known, this would enable solving any $\mathcal{NP}$ decision problem efficiently by reduction, including the challenging 3-SAT problem. The relations between the different complexity classes are illustrated in Figure 2.6.

Finally, the difficulty of an optimization problem is measured in terms of a corresponding decision problem. If the decision problem is already too difficult to decide, the optimization will be even more challenging to solve. For example, let $\min_{x \in X} f(x)$ denote some optimization problem for some set X and function f. We obtain a corresponding decision problem by deciding for $c \in \mathbb{R}$ whether an input $x \in X$ exists such that $f(x) \leq c$. It is thus common to say that an optimization problem is in $\mathcal{NP}$ or $\mathcal{NP}$-hard if the corresponding decision problem is in $\mathcal{NP}$ or $\mathcal{NP}$-hard.

The tracking method of Chapter 3 utilizes an optimization problem that is $\mathcal{NP}$-hard, as shown in Section 3.2.2.1 by reduction to a binary linear optimization problem. Nevertheless, to obtain high-quality approximations, an efficient solver that provides long-term consistent trajectories is presented in Chapter 3.

The tracking formulation presented in Chapter 4 results in an $\mathcal{NP}$-hard problem as well. To prove this, a special case of the formulation is considered that is shown to be reducible to the 3-SAT problem in Section 4.6. Yet, this work presents an algorithm that enables to obtain a global optimal solution in practice.

---

[4] The image has been adapted from the source [159].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

2.6 Optimization Theory

All MOT methods presented in the subsequent chapters obtain their trajectories by solving a difficult optimization problem. The methods of Chapter 3 employ a binary quadratic optimization problem, while a binary linear optimization problem is used in Chapter 4. The underlying theory for these solvers is briefly introduced in this section.

2.6.1 Linear programming

Linear programming deals with optimization problems[5] of the form $\min_{x \in \mathrm{P}} f(x)$, where the objective function f is a linear functional over $\mathbb{R}$ and P describes the constraints in terms of finitely many linear inequalities. A good understanding of the theory has led to efficient solvers that can tackle huge[6] problem instances. We give a short introduction to the theory, following the notation of Grötschel [156], and refer the interested reader to standard textbooks [156–158] for further information.

In linear programming, it is common to generalize the relations '$\leq$','$\geq$' and '$=$' to vectors, *e.g.*, given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we set

$$\mathbf{x} \geq \mathbf{y} :\Longleftrightarrow x_i \geq y_i, \forall i \in [n]\,. \tag{2.24}$$

The other relations are defined accordingly.

Definition 2.37. For $n \in \mathbb{N}$, $\mathbf{g} \in \mathbb{R}^n$ and $y \in \mathbb{R}$, the set $H \coloneqq \{\mathbf{x} \in \mathbb{R}^n \mid \langle \mathbf{g}, \mathbf{x} \rangle = y\}$ is called *hyperplane* and $H^{\leq} \coloneqq \{\mathbf{x} \in \mathbb{R}^n \mid \langle \mathbf{g}, \mathbf{x} \rangle \leq y\}$ (closed) *half-space*.

The next definition introduces the central objects in linear programming.

Definition 2.38. A convex *polyhedron* $\mathrm{P} \subseteq \mathbb{R}^n$ is a finite intersection of closed half-spaces. P is in *canonical $\mathcal{H}$-representation*, denoted as $\mathrm{P}(\mathbf{A}, \mathbf{b}) \coloneqq \mathrm{P}$, if for a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and vector $\mathbf{b} \in \mathbb{R}^m$, $\mathrm{P} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$.

We first note that a canonical form may encode upper bounded and equality constraints: If $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, we have

$$\langle \mathbf{a}, \mathbf{x} \rangle \leq b \Longleftrightarrow -\langle \mathbf{a}, \mathbf{x} \rangle \geq -b, \text{ and } \langle \mathbf{a}, \mathbf{x} \rangle = b \Longleftrightarrow \begin{pmatrix} \mathbf{a}^\intercal \\ -\mathbf{a}^\intercal \end{pmatrix} \mathbf{x} \geq \begin{pmatrix} b \\ -b \end{pmatrix}. \tag{2.25}$$

Note that any set $\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}\}$ can be transformed into canonical representation by setting $\mathbf{x} = \mathbf{x}_1 - \mathbf{x}_2$ for suitable $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n_{\geq 0}$. We conclude that any polyhedron can be written in canonical form, indeed.

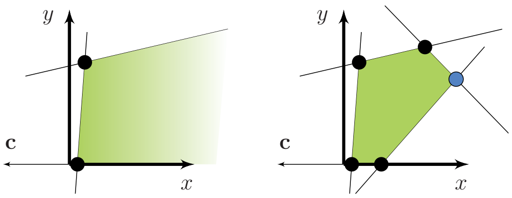
The goal of linear programming is to solve linear optimization problems.

---

[5] We will focus on minimization problems. The statements are adapted for maximization problems accordingly.

[6] On a state-of-the-art benchmark [155] for linear programming, problem instances have up to $1.6 \cdot 10^6$ many variables and up to $1.1 \cdot 10^6$ many constraints.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a) (b)

Figure 2.7: Depicted is the cost vector $\mathbf{c} = (-1, 0)^\intercal$ and two polyhedrons. Circles mark the vertices of the respective polyhedron. (a) The colored area shows an unbounded polyhedron with no optimal (minimal) solution. (b) The bounded polyhedron has a unique optimal (minimal) solution given by the position of the blue circle.

Definition 2.39. Let $\mathrm{P} \subseteq \mathbb{R}^n$ be a polyhedron and $\mathbf{c} \in \mathbb{R}^n$. A *linear program* (LP) defined by $(\mathrm{P}, \mathbf{c})$ is the task to compute

$$\eta^{(\mathrm{P},\mathbf{c})} := \inf\{\langle \mathbf{c}, \mathbf{x} \rangle \mid \mathbf{x} \in \mathrm{P}\}\,. \tag{2.26}$$

In this context, P is called the *feasibility set*. Each vector $\mathbf{x} \in \mathrm{P}$ is a *valid solution* and $\mathbf{c}$ is the *objective function*. Any vector $\mathbf{x}^\star \in \mathrm{P}$ satisfying

$$\langle \mathbf{c}, \mathbf{x}^\star \rangle = \min\{\langle \mathbf{c}, \mathbf{x} \rangle \mid \mathbf{x} \in \mathrm{P}\} = \eta^{(\mathrm{P},\mathbf{c})} \tag{2.27}$$

is an *optimal solution* and $\langle \mathbf{c}, \mathbf{x}^\star \rangle$ is the corresponding *optimal value*.

The question arises under which conditions a *Linear Program* (LP) has an optimal solution and how to compute it. The following theorem addresses the first question:

Theorem 2.40. *An LP defined by* $(\mathrm{P}, \mathbf{c})$ *contains an optimal solution if and only if* P *contains a valid solution and* $\langle \mathbf{c}, \cdot \rangle$ *is lower bounded in* P.

Proof. See Grötschel [156]. □

Exemplary polyhedrons with no resp. one optimal solution are shown in Figure 2.7. In order to characterize valid solutions, the vertices of a polyhedron play a key role. Given a polyhedron P, a point $\mathbf{p} \in \mathrm{P}$ is a *vertex* of P if and only if, for all $\mathbf{p}_1, \mathbf{p}_2 \in \mathrm{P}$ and $\lambda \in [0, 1]$ with $\mathbf{p} = \lambda \mathbf{p}_1 + (1 - \lambda)\mathbf{p}_2$, either $\mathbf{p} = \mathbf{p}_1$ or $\mathbf{p} = \mathbf{p}_2$ holds.

The following theorem shows the relation between a valid solution, an optimal solution, and the vertices of the corresponding polyhedron.

Theorem 2.41. *Let* $(\mathrm{P}, \mathbf{c})$ *be an LP. Then the following holds:*

(i) P *has only finitely many (possibly zero) vertices.*

(ii) If $(\mathrm{P}, \mathbf{c})$ *contains an optimal solution, then its set of optimal solutions contains a vertex of* P.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Proof. See Grötschel [156]. ☐

Thus, assuming that we already know that the objective function $\mathbf{c}$ is bounded in P, a naïve algorithm to obtain an optimal solution is to evaluate $\langle \mathbf{c}, \mathbf{x} \rangle$ for all vertices $\mathbf{x}$ of P. As a polyhedron can have exponentially many vertices with respect to the dimension of P, such an approach is not tractable in practice, though.

The simplex algorithm [160] follows the same principle of checking vertices but uses sophisticated rules to define a traversal order. The algorithm allows deciding whether the current solution (vertex) is already optimal or whether no optimal solution exists without looping over all vertices[7].

2.6.2 Binary linear programming

A modification of a linear program is a BLP. It requires that the feasibility set of a linear program contains only binary vectors. This allows the formulation of many discrete optimization problems using BLPs and is used in particular in the tracking formulation described in Chapter 4 to obtain the correct associations of detections to consistent trajectories.

Definition 2.42. For a polyhedron $\mathrm{P} \subseteq \mathbb{R}^n$, let $\mathrm{P}_\mathcal{B} := \mathrm{P} \cap \{0,1\}^n$. A *binary linear program* (BLP) defined by $(\mathrm{P}_\mathcal{B}, \mathbf{c})$ is the task to compute

$$\eta^{(\mathrm{P}_\mathcal{B}, \mathbf{c})} = \inf \{ \langle \mathbf{c}, \mathbf{x} \rangle \mid \mathbf{x} \in \mathrm{P}_\mathcal{B} \} . \tag{2.28}$$

For $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, we set $\mathrm{P}_\mathcal{B}(\mathbf{A}, \mathbf{b}) := \mathrm{P}(\mathbf{A}, \mathbf{b}) \cap \{0,1\}^n$.

Since $\mathrm{P}_\mathcal{B} \subseteq \{0,1\}^n$, any BLP contains at most finitely many valid solutions. An exhaustive search thus provides the optimal solution if there is any, but such an approach is not tractable for all BLPs. Due to the following theorem, no algorithm with a better time complexity is known, when applied to arbitrary BLPs:

Theorem 2.43. *For an arbitrary BLP, computing an optimal solution is an $\mathcal{NP}$-complete problem.*

Proof. See Grötschel [156]. ☐

Still, many BLP problem instances can be solved to optimality nowadays. The subsequent part of this section shows different approaches to tackle BLPs.

2.6.2.1 Relaxations.

Before discussing approaches to obtain optimal solutions, we note that it is sometimes sufficient to approximate an optimal solution of a BLP using a *relaxation*: Given a BLP defined by $(\mathrm{P}_\mathcal{B}, \mathbf{c})$ with $\mathrm{P}_\mathcal{B} \subseteq \{0,1\}^n$, the optimization is performed on a feasibility set

---

[7] There are still problem instances with exponentially many vertices requiring the simplex algorithm to loop over all vertices before terminating. Yet, the expected runtime is polynomial.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

$\mathrm{P}' \supseteq \mathrm{P}_{\mathcal{B}}$ by ignoring some constraints of $\mathrm{P}_{\mathcal{B}}$. The set $\mathrm{P}' \subset \mathbb{R}^n$ is either a polyhedron or the intersection of a polyhedron with $\{0,1\}^n$. Finally, a solution $\mathbf{x}$ of P' must be projected to a valid solution of $\mathrm{P}_{\mathcal{B}}$. A commonly used relaxation is the LP-relaxation:

Definition 2.44. Let $\mathrm{P} \subseteq \mathbb{R}^n$ be a polyhedron with corresponding set $\mathrm{P}_{\mathcal{B}}$. The *LP-relaxation* of $\mathrm{P}_{\mathcal{B}}$ is given by $\mathrm{P}^{\circ}_{\mathcal{B}} \coloneqq \mathrm{P} \cap [0,1]^n$. Thus, the LP-relaxation removes the binary constraints, allowing each component to be within $[0,1]$. In general, the *continuous relaxation* of an optimization problem replaces each binary constraint with a $[0,1]$ box constraint.

Since the LP-relaxation is an LP, standard linear programming can be applied to solve the optimization problem on the relaxation efficiently. For certain types of polyhedrons, any optimal solution of the LP-relaxation is an optimal solution of the corresponding BLP. A sufficient condition is given by total unimodularity.

Definition 2.45. A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is called *totally unimodular* if, for all quadratic submatrices $\mathbf{A}'$ of $\mathbf{A}$, we have $\det(\mathbf{A}') \in \{0, 1, -1\}$.

The next theorem connects BLPs with the total unimodularity property.

Theorem 2.46. *If $\mathbf{A} \in \mathbb{R}^{m \times n}$ is totally unimodular and $\mathbf{b} \in \mathbb{Z}^m$, then all vertices of $\mathrm{P}(\mathbf{A}, \mathbf{b})$ are integral.*

Proof. See Schrijver [157]. ☐

As a consequence of Theorem 2.46, applying the simplex algorithm on the LP-relaxation of a BLP $(\mathrm{P}_{\mathcal{B}}(\mathbf{A}, \mathbf{b}), \mathbf{c})$ for a totally unimodular matrix $\mathbf{A}$ and $\mathbf{b} \in \mathbb{Z}^m$ returns the solution to the BLP $(\mathrm{P}_{\mathcal{B}}(\mathbf{A}, \mathbf{b}), \mathbf{c})$. The solver presented in Chapter 3 takes advantage of this theorem.

2.6.2.2 Optimal solutions.

Branch-and-bound. If $\mathbf{A}$ is not totally unimodular, an optimal solution can still be obtained, if it exists, using multiple LP-relaxations according to the branch-and-bound algorithm [158, 161].

Let $(\widetilde{\mathrm{P}}_{\mathcal{B}}, \mathbf{c})$ be a BLP. At each iteration of the branch-and-bound algorithm, a set K of polyhedrons is maintained. Solving the BLP $(\mathrm{P}_{\mathcal{B}}, \mathbf{c})$, for all $\mathrm{P} \in K$, will result in an optimal binary solution of the BLP $(\widetilde{\mathrm{P}}_{\mathcal{B}}, \mathbf{c})$, if it exists. Let $\eta^{\star}_{\mathcal{B}}$ denote the best-known objective value for the binary optimization problem $(\widetilde{\mathrm{P}}_{\mathcal{B}}, \mathbf{c})$ so far.

Initially, $K = \{\widetilde{\mathrm{P}}^{\circ}_{\mathcal{B}}\}$, where $\widetilde{\mathrm{P}}^{\circ}_{\mathcal{B}}$ is the LP-relaxation of $\widetilde{\mathrm{P}}_{\mathcal{B}}$. Within a loop, a polyhedron $\mathrm{P} \in K$ is extracted and the simplex algorithm is applied on P which results in three cases:

1. If $\mathrm{P} = \emptyset$ is returned, the polyhedron is ignored and the algorithm continues with the next polyhedron $\mathrm{P} \in K$ if available.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

2. The simplex algorithm returns an optimal solution $\mathbf{x}_{\text{opt}}$ that is not binary. Let η be the corresponding optimal value. If $\eta \geq \eta^\star_{\mathcal{B}}$, P can be ignored as a better bound on the optimal value is already known. If $\eta < \eta^\star_{\mathcal{B}}$, the polyhedron P may contain a valid binary solution of $\widetilde{\mathrm{P}}_{\mathcal{B}}$ with a lower objective value than $\eta^\star_{\mathcal{B}}$. Therefore, P is decomposed into $\mathrm{P} = \mathrm{P}^{(1)} \sqcup \ldots \sqcup \mathrm{P}^{(k)}$ for polyhedrons $\mathrm{P}^{(1)}, \ldots, \mathrm{P}^{(k)}$. This step is called *branching*. The relation between the optimal value of P and a polyhedron $\mathrm{P}^{(i)}$, for all $i \in [k]$, is given by

$$\eta^{(\mathrm{P}^{(i)}_{\mathcal{B}}, \mathbf{c})} \geq \eta^{(\mathrm{P}^{(i)}, \mathbf{c})} \geq \eta^{(\mathrm{P}, \mathbf{c})} \,. \tag{2.29}$$

   An exemplary branching step is to decompose P into $\mathrm{P}^{(1)} = \{\mathbf{x} \in \mathrm{P} \mid x_1 \leq 0.5\}$ and $\mathrm{P}^{(2)} = \{\mathbf{x} \in \mathrm{P} \mid x_1 \geq 0.5\}$ which means that each binary solution $\mathbf{x}$ in $\mathrm{P}^{(1)}$ satisfies $x_1 = 0$, while each binary solution in $\mathrm{P}^{(2)}$ satisfies $x_1 = 1$. Finally, the polyhedrons $\mathrm{P}^{(1)}, \ldots, \mathrm{P}^{(k)}$ replace P in K.

3. The simplex algorithm returns an optimal binary solution $\mathbf{x}_{\text{opt}}$ for P with optimal value $\eta_{\mathcal{B}}$. If $\eta_{\mathcal{B}} < \eta^\star_{\mathcal{B}}$, a better binary solution has been found which creates a new bound $\eta^\star_{\mathcal{B}}$. Accordingly, all polyhedrons $\mathrm{P} \in K$ with optimal value $\eta^{(\mathrm{P}, \mathbf{c})} \geq \eta^\star_{\mathcal{B}}$ are removed. In particular, a polyhedron P' can be removed from K if it stems from the branching step of a polyhedron P with $\eta^{(\mathrm{P}, \mathbf{c})} \geq \eta^\star_{\mathcal{B}}$, according to Eq. (2.29).

A pseudo-code of the algorithm is provided in Algorithm 2.3.

Branch-and-cut. Besides the branch-and-bound algorithm, an alternative procedure is to iteratively remove non-binary vertices of the convex hull of a BLP. The following theorem provides the basis for this algorithm.

Theorem 2.47. *The optimal value of a BLP* $(\mathrm{P}_{\mathcal{B}}, \mathbf{c})$ *can be computed on its convex hull,* i.e.*,*

$$\inf \{\langle \mathbf{c}, \mathbf{x} \rangle \mid \mathbf{x} \in \mathrm{P}_{\mathcal{B}}\} = \inf \{\langle \mathbf{c}, \mathbf{x} \rangle \mid \mathbf{x} \in \operatorname{conv}(\mathrm{P}_{\mathcal{B}})\} \,. \tag{2.30}$$

Furthermore, $\operatorname{conv}(\mathrm{P}_{\mathcal{B}})$ *is a (bounded) polyhedron.*

Proof. See Grötschel [156]. ☐

Therefore, a BLP $(\mathrm{P}_{\mathcal{B}}, \mathbf{c})$, *i.e.*, the left side of Eq. (2.30), can be solved by formulating it as an LP (right side of Eq. (2.30)). While Theorem 2.47 explains how to solve a BLP in theory, it is very difficult in practice to obtain $\operatorname{conv}(\mathrm{P}_{\mathcal{B}})$. In the following, we fix an objective function $\mathbf{c}$ and the LP-relaxation of $\mathrm{P}_{\mathcal{B}}$. The *cutting-plane algorithm* (*e.g.*, see Bertsimas *et al.* [158]) leverages Theorem 2.47 to tackle BLPs effectively in practice.

The cutting-plane algorithm iteratively downsizes the feasibility set by introducing additional constraints in the form of closed half-spaces. At iteration t, the feasibility set is given by $\mathrm{P}_t \coloneqq \mathrm{P} \cap \bigcap_{i=1}^{t-1} H_i^{\leq}$, where each H_i is a hyperplane. If P_t results in an optimal (vertex) solution $\mathbf{x}_{\text{opt}}$ that is not binary, the solution is excluded in the next polyhedron P_{t+1}. To this end, the cutting-plane algorithm uses a new hyperplane H_t such that $\mathbf{x}_{\text{opt}} \notin H_t^{\leq}$, while all binary valid solutions remain still valid in P_{t+1}, *i.e.*, $\mathrm{P}_{\mathcal{B}} \subset \mathrm{P}_{t+1} \coloneqq \mathrm{P}_t \cap H_t^{\leq}$.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Algorithm 2.3: Branch-and-bound

```
Input  : BLP defined by $(\mathrm{P}_{\mathcal{B}}, \mathbf{c})$.
Output : Optimal binary solution $\mathbf{x}^\star$ with optimal value $\eta^\star_{\mathcal{B}}$.
$K \leftarrow \{\mathrm{P}\}$
$\eta^\star_{\mathcal{B}} \leftarrow \infty$
$\mathbf{x}^\star \leftarrow$ null
while $K \neq \emptyset$ do
    Select polyhedron $\mathrm{P} \in K$
    $K \leftarrow K \setminus \{\mathrm{P}\}$
    if $\mathrm{P} \neq \emptyset$ then
        $\mathbf{x}_{\text{opt}} \leftarrow \arg\min_{\mathbf{x} \in \mathrm{P}} \langle \mathbf{c}, \mathbf{x} \rangle$
        $\eta^{(\mathrm{P},\mathbf{c})} \leftarrow \langle \mathbf{c}, \mathbf{x}_{\text{opt}} \rangle$
        if $\eta^{(\mathrm{P},\mathbf{c})} < \eta^\star_{\mathcal{B}}$ then
            if $\mathbf{x}_{\text{opt}}$ binary then
                $\mathbf{x}^\star \leftarrow \mathbf{x}_{\text{opt}}$
                $\eta^\star_{\mathcal{B}} \leftarrow \eta^{(\mathrm{P},\mathbf{c})}$
                Remove all $\mathrm{P}' \in K$ from $K$ with optimal value $\eta^{(\mathrm{P}',\mathbf{c})} \geq \eta^\star_{\mathcal{B}}$.
            end if
            if $\mathbf{x}_{\text{opt}}$ not binary then
                Branch P into $\mathrm{P}_1, \ldots, \mathrm{P}_k$.
                $K \leftarrow K \cup \{\mathrm{P}_1, \ldots, \mathrm{P}_k\}$
            end if
        end if
    end if
end while
```

The half-space $H_t^{\leq}$ is called *cutting hyperplane* or *cut constraint.* The goal of the cutting-plane algorithm is to shrink P_t such that for each iteration t the chain

$$\operatorname{conv}(\mathrm{P}_{\mathcal{B}}) \subseteq \mathrm{P}_t \subsetneq \mathrm{P}_{t-1} \subsetneq \ldots \subsetneq \mathrm{P}_1 \subsetneq \mathrm{P} \tag{2.31}$$

holds. Conceptually, the hope is that eventually at some iteration t, P_t is so close to $\operatorname{conv}(\mathrm{P}_{\mathcal{B}})$ that its optimal solution is binary. The difficulty in using the cutting-plane method thus lies in efficiently finding new cut constraints. Several cut constraints are known that can be used for BLPs (*e.g.*, Gomory cuts [162]).

Finally, if no new cut-constraint can be found, the branch-and-bound algorithm can be applied to the remaining polyhedron. Combining branch-and-bound with the cutting-plane algorithm is called *branch-and-cut* [163, 164]. A pseudo-code of the branch-and-cut algorithm is provided in Algorithm 2.4 and an illustration of the cutting plane algorithm in Figure 2.8.

The proposed solvers in this work build upon the software Gurobi [165], which is an optimal BLP solver that implements the branch-and-bound and branch-and-cut algorithms.

Separation algorithm on $\mathrm{P}_{\mathcal{B}}$. The cutting-plane and branch-and-bound algorithms start with the LP-relaxation of $\mathrm{P}_{\mathcal{B}}$. However, this is not practicable if the problem has

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Algorithm 2.4: Branch-and-cut

```
   Input   : BLP defined by $(\mathrm{P}_{\mathcal{B}}, \mathbf{c})$.
   Output  : Optimal binary solution $\mathbf{x}^\star$ with optimal value $\eta^\star_{\mathcal{B}}$.
 1 $K \leftarrow \{\mathrm{P}\}$
 2 $\eta^\star_{\mathcal{B}} \leftarrow \infty$
 3 $\mathbf{x}^\star \leftarrow$ null
 4 while $K \neq \emptyset$ do
 5     Select polyhedron $\mathrm{P} \in K$
 6     $K \leftarrow K \setminus \{\mathrm{P}\}$
 7     if $\mathrm{P} \neq \emptyset$ then
 8         $\mathbf{x}_{\mathrm{opt}} \leftarrow \arg\min_{\mathbf{x} \in \mathrm{P}} \langle \mathbf{c}, \mathbf{x} \rangle$
 9         $\eta^{(\mathrm{P},\mathbf{c})} \leftarrow \langle \mathbf{c}, \mathbf{x}_{\mathrm{opt}} \rangle$
10         if $\eta^{(\mathrm{P},\mathbf{c})} < \eta^\star_{\mathcal{B}}$ then
11             if $\mathbf{x}_{\mathrm{opt}}$ binary then
12                 $\mathbf{x}^\star \leftarrow \mathbf{x}_{\mathrm{opt}}$
13                 $\eta^\star_{\mathcal{B}} \leftarrow \eta^{(\mathrm{P},\mathbf{c})}$
14                 Remove all $\mathrm{P}' \in K$ from $K$ with optimal value $\eta^{(\mathrm{P}',\mathbf{c})} \geq \eta^\star_{\mathcal{B}}$.
15                 Goto line 4
16             else
17                 if cut constraint $H$ available then
18                     $\mathrm{P} \leftarrow \mathrm{P} \cap H^{\leq}$
19                     Goto line 7
20                 end if
21                 Branch P into polyhedrons $\mathrm{P}_1, \ldots, \mathrm{P}_k$.
22                 $K \leftarrow K \cup \{\mathrm{P}_1, \ldots, \mathrm{P}_k\}$
23             end if
24         end if
25     end if
26 end while
```

an exponential growth in the number of constraints.

An alternative approach is to start with a BLP-relaxation of $\mathrm{P}_{\mathcal{B}}$. In an iterative process, constraints of $\mathrm{P}_{\mathcal{B}}$ are added that are violated by the current solution and the resulting BLP is solved again. It is thus a variant of the cutting-plane algorithm that is based on the separation problem: Given a polyhedron $\mathrm{P} \subseteq \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^n$, the *separation problem* is to decide whether $\mathbf{x}$ satisfies all constraints of P (thus if $\mathbf{x} \in \mathrm{P}$) or if not, output a cutting hyperplane that represents constraints violated by $\mathbf{x}$. Such a function is also called *oracle*[8]. An efficient oracle function within a separation algorithm is used in Chapter 4 to handle the BLP of the presented HO-MOT formulation despite the exponentially many constraints. Pseudo-code of the algorithm is provided in Algorithm 2.5.

---

[8] An oracle with a polynomial runtime leads to an LP solver with polynomial runtime. The ellipsoid method is such an oracle, leading to polynomial runtime for solving LPs [166].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

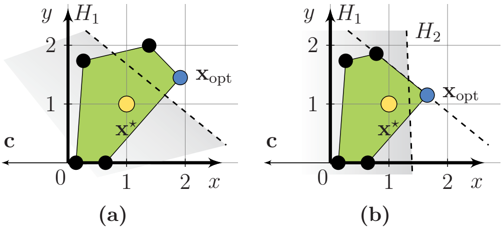


Figure 2.8: Illustration of the cutting-plane algorithm (for minimization), with $\mathbf{c} = (-1, 0)^\intercal$. (a) The polyhedron (green area) contains exactly one binary solution (yellow circle). A cutting hyperplane H_1 separates the optimal LP solution $\mathbf{x}_{\text{opt}}$ (blue circle) from the binary solution. The closed half-space is the gray area. (b) In the next iteration, the LP is solved on the intersection with H_1, and the next separating hyperplane H_2 is computed.

Algorithm 2.5: Separation algorithm

```
Input  : BLP defined by $(\mathrm{P}_\mathcal{B}, \mathbf{c})$.
Output : Optimal solution $\mathbf{x}^\star$.
/* Return a relaxation of $\mathrm{P}_\mathcal{B}$.                                  */
$\mathrm{P}'_\mathcal{B} \leftarrow$ Relaxation($\mathrm{P}_\mathcal{B}$)
do
    $\mathbf{x}^\star \leftarrow \arg\min_{\mathbf{x}\in\mathrm{P}'_\mathcal{B}} \langle \mathbf{c}, \mathbf{x} \rangle$
    /* Return a cutting plane $H^\leq \neq \emptyset$ if it exists, and $\emptyset$ otherwise.   */
    $H^\leq \leftarrow$ Oracle($\mathrm{P}_\mathcal{B}, \mathbf{x}^\star$)
    if $H^\leq \neq \emptyset$ then
        $\mathrm{P}'_\mathcal{B} \leftarrow \mathrm{P}'_\mathcal{B} \cap H^\leq$
    end if
while $H \neq \emptyset$
```

2.6.3 Quadratic programming

The cost function $\mathbf{c} = (c_j)_{j\in[n]} \in \mathbb{R}^n$ of a linear program assigns each component x_i of a valid solution $\mathbf{x} = (x_j)_{j\in[n]} \in \mathbb{R}^n$ the value $c_i x_i$. Yet, it is often desirable to have a cost function that assigns in addition a value $q_{i,j} \in \mathbb{R}$ depending on the variables x_i and x_j, resulting in $q_{i,j}x_i x_j + c_i x_i + c_j x_j$. This is formalized in a quadratic program:

Definition 2.48. Let $\mathrm{P} \subseteq \mathbb{R}^n$ be a polyhedron, $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{Q} \in \mathbb{R}^{n\times n}$. A *Quadratic Program* (QP) defines the optimization problem

$$\eta^{(\mathrm{P},\mathbf{Q},\mathbf{c})} = \inf\left\{ \frac{1}{2}\mathbf{x}^\intercal\mathbf{Q}\mathbf{x} + \mathbf{c}^\intercal\mathbf{x} \mid \mathbf{x} \in P \right\} . \tag{2.32}$$

If $\mathbf{Q}$ is positive-definite, problem (2.32) can be solved in polynomial time [167] using convex optimization techniques [167] such as the Frank-Wolfe algorithm [117]. In fact, the Frank-Wolfe algorithm can be applied to any convex and differentiable function $f : \mathrm{P} \to \mathbb{R}$ on a convex set P. The algorithm considers at the k-th iteration the best known valid solution for f, which we denote by $\mathbf{x}^{(k)}$. Then, the function f is

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

approximated by the first-order Taylor polynomial $\tilde{f}$ around $\mathbf{x}^{(k)}$. Thus instead of minimizing f, an optimal solution $\mathbf{a}^{(k)}$ for function $\tilde{f}$ within P is computed. Finally, the next iterate $\mathbf{x}^{(k+1)}$ is chosen to be a vector on the line spanned by $\mathbf{a}^{(k)}$ and $\mathbf{x}^{(k)}$ that minimizes f. A pseudo-code of the algorithm is provided in Algorithm 2.6. Yet, if $\mathbf{Q}$ is indefinite or has at least one negative eigenvalue, the QP problem is $\mathcal{NP}$-hard [168].

Algorithm 2.6: Frank-Wolfe

```
Input  : Polyhedron P with feasible solution $\mathbf{x}^0$ and objective function $f$.
Output : Optimal solution $\mathbf{x}^\star$.
$k \leftarrow 0$
while Stopping criteria not satisfied do
    $\mathbf{a}^{(k)} \leftarrow \arg\min_{\mathbf{a}\in \mathrm{P}} \nabla f(\mathbf{x}^{(k)})\mathbf{a}$
    $\eta^\star \leftarrow \arg\min_{0\leq\eta\leq 1} f(\mathbf{x}^{(k)} + \eta(\mathbf{a}^{(k)} - \mathbf{x}^{(k)}))$
    $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + \eta^\star(\mathbf{a}^{(k)} - \mathbf{x}^{(k)})$
    $k \leftarrow k + 1$
end while
```

2.6.4 Binary quadratic programming

Particularly important for this work are QP problems that require valid solutions to be binary.

Definition 2.49. Let P be a polyhedron, $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{Q} \in \mathbb{R}^{n\times n}$. A *Binary Quadratic Program* (BQP), denoted by $(\mathrm{P}_\mathcal{B}, \mathbf{Q}, \mathbf{c})$, defines the optimization problem

$$\eta^{(\mathrm{P}_\mathcal{B},\mathbf{Q},\mathbf{c})} := \inf\left\{\frac{1}{2}\mathbf{x}^\intercal\mathbf{Q}\mathbf{x} + \mathbf{c}^\intercal\mathbf{x} \mid \mathbf{x} \in \mathrm{P}_\mathcal{B}\right\}. \tag{2.33}$$

Like for BLPs, solving a BQP poses an $\mathcal{NP}$-hard problem, which follows immediately from Theorem 2.43 by setting $\mathbf{Q} = \mathbf{0}$. Note also that each BQP can be transformed into an equivalent BLP by the following theorem:

Theorem 2.50. *Each BQP can be transformed into a BLP without changing the set of optimal solutions.*

Proof. Let $(\mathrm{P}_\mathcal{B}, \mathbf{Q}, \mathbf{c})$ be a BQP in $\mathbb{R}^n$ as in Definition 2.49. For $\mathbf{x} = (x_i)_{i\in[n]} \in \mathrm{P}_\mathcal{B}$, let

$$\widetilde{\mathrm{P}}(\mathbf{x}) := \left\{\mathbf{Z} = (z_{i,j})_{i,j\in[n]} \in \{0,1\}^{n\times n} \;\middle|\; \begin{array}{l} z_{i,j} \leq x_i,\ \forall i,j \in [n], \\ z_{i,j} \leq x_j,\ \forall i,j \in [n], \\ z_{i,j} \geq x_i + x_j - 1,\ \forall i,j \in [n] \end{array}\right\}. \tag{2.34}$$

Given $\mathbf{A} \in \mathbb{R}^{n\times n}$, we define the column vector $\mathrm{vec}(\mathbf{A}) \in \mathbb{R}^{n^2}$ by stacking the columns of $\mathbf{A}$. For $\mathbf{x} \in \mathrm{P}_\mathcal{B}$ and $\mathbf{Z} \in \widetilde{\mathrm{P}}(\mathbf{x})$, since $z_{i,j} = x_i x_j$ for all $i, j \in [n]$, it follows that $|\widetilde{\mathrm{P}}(\mathbf{x})| = 1$. Hence, each $\mathbf{x} \in \mathrm{P}_\mathcal{B}$ uniquely defines a vector $\mathbf{z}(\mathbf{x})$ such that $(\mathbf{z}(\mathbf{x}))^\intercal = \mathrm{vec}(\mathbf{Z})$ and $\mathbf{Z} \in \widetilde{\mathrm{P}}(\mathbf{x})$. With $\mathbf{q}' := \mathrm{vec}(\frac{1}{2}\mathbf{Q}) \in \mathbb{R}^{n^2}$, $\mathrm{P}'_\mathcal{B} := \{(\mathbf{z}(\mathbf{x}), \mathbf{x}^\intercal)^\intercal \mid \mathbf{x} \in \mathrm{P}_\mathcal{B}\} \subset \mathbb{R}^{n^2+n}$ and objective function $\mathbf{c}' := \begin{pmatrix}\mathbf{q}'^\intercal & \mathbf{c}^\intercal\end{pmatrix}^\intercal \in \mathbb{R}^{n^2+n}$, the BLP $(\mathrm{P}'_\mathcal{B}, \mathbf{c}')$ preserves the optimal solutions of $\mathrm{P}_\mathcal{B}$, since $\frac{1}{2}\mathbf{x}^\intercal\mathbf{Q}\mathbf{x} + \mathbf{c}^\intercal\mathbf{x} = \mathbf{c}'^\intercal\mathbf{p}(\mathbf{x})$ for $\mathbf{p}(\mathbf{x}) = (\mathbf{z}(\mathbf{x}), \mathbf{x}^\intercal)^\intercal \in \mathrm{P}'_\mathcal{B}$.

□

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Chapter 3 formulates implicit HO-MOT as a binary quadratic problem such that the consistency between all detections of a trajectory is ensured. In order to tackle the corresponding BQP, an efficient approximate solver derived from the Frank-Wolfe algorithm is presented in Section 3.3.

2.6.5 Non-linear optimization

Let $F : \mathbb{R}^m \to \mathbb{R}^n$ be differentiable and $m < n$. A problem that appears frequently in computer vision is to compute

$$\min_{\mathbf{x} \in \mathbb{R}^m} \frac{1}{2} ||F(\mathbf{x})||^2. \tag{2.35}$$

If $\mathbf{x}^{(0)} \in \mathbb{R}^m$ is an initial guess of the solution, we obtain an approximate solution using the Levenberg-Marquardt algorithm by iterating via

$$\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} - \alpha_k \left(\mathbf{D}(\mathbf{x}^{(k)})^\intercal \mathbf{D}(\mathbf{x}^{(k)}) + \lambda_k \mathbf{I} \right)^{-1} \mathbf{D}(\mathbf{x}^{(k)})^\intercal F(\mathbf{x}^{(k)}), \tag{2.36}$$

where $\mathbf{D}(\mathbf{x}^{(k)})$ is the differential of F at $\mathbf{x}^{(k)}$, $\mathbf{I} \in \mathbb{R}^{m \times m}$ is the identity matrix, and $\alpha_k \in \mathbb{R}_{\geq 0}$ is the step size. The value $\lambda_k \in \mathbb{R}_{\geq 0}$ is chosen such that

$$\mathbf{M}_k := \left(\mathbf{D}(x^{(k)})^\intercal \mathbf{D}(x^{(k)}) + \lambda_k \mathbf{I} \right)^{-1} \tag{2.37}$$

is positive-definite. For $f(\mathbf{x}) := \frac{1}{2}||F(\mathbf{x})||^2$, we have $\nabla f(\mathbf{x}) = \mathbf{D}(\mathbf{x})^\intercal F(\mathbf{x})$ so that $d_k := -\mathbf{M}_k \mathbf{D}(\mathbf{x}^{(k)})^\intercal F(\mathbf{x}^{(k)})$ is a descent direction, according to Theorem 2.20, which motivates the iteration procedure (2.36). For large λ_k, Eq. (2.36) can be considered as a gradient descent of F. For $\lambda_k = 0$, Eq. (2.36) corresponds to an iteration of the Gauß-Newton algorithm. The Levenberg-Marquardt algorithm is used in Section 3.5.2.4 to reconstruct missing detections.

2.7 Multi-Object Tracking

While the mathematical fundamentals that have been introduced so far enable us to understand challenges and derive solutions for the data association, this section introduces the remaining aspects of a multiple object tracking method.

All MOT methods proposed in this work employ off-the-shelf object detectors that are also used by compared tracking methods to enable meaningful comparisons. Section 2.7.1 briefly introduces the corresponding object detectors. Two types of pairwise features are introduced in Section 2.7.2 as they build the basis for the MOT methods proposed in this work. State-of-the-art MOT datasets are introduced in Section 2.7.3. They are used to assess the proposed tracking methods. Finally, Section 2.7.4 introduces state-of-the-art MOT metrics to measure the quality of a tracking result.

2.7.1 Object detectors

MOT methods based on the tracking-by-detection paradigm rely to a great extent on the quality of the input detections. As this thesis focuses on improving the data

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

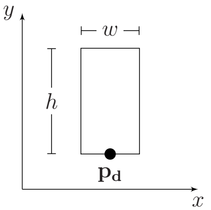


Figure 2.9: A detection $\mathbf{d}$ is described by its width w, height h, and image coordinates $\mathbf{p_d}$, as depicted in the figure.

association part, detections are retrieved using off-the-shelf object detectors. This allows comparing among different trackers the improvement due to the data association. The aim of an object detector is to locate any object of interest visible in a video in the form of a corresponding detection defined as follows:

Definition 2.51. A *detection* $\mathbf{d}$ is an axis-aligned bounding box defining the location of an object in a particular image. A detection can be represented as $\mathbf{d} = (x, y, w, h, \mathfrak{f}, p) \in \mathbb{R}^6$. If not otherwise stated, $\mathbf{p_d} \coloneqq (x, y)$ corresponds to the center position of the lower edge of the bounding box in image coordinates, which we call the *lower central anchor point* of $\mathbf{d}$. The values w and h measure the width and height of the bounding box in image coordinates, respectively (see also Figure 2.9). The value $\mathfrak{f}$ denotes the timestamp (a frame number) of the corresponding image. Associated to each detection $\mathbf{d}$ is a confidence $p \in [0, 1]$ indicating the likelihood of $\mathbf{d}$ being a correct detection.

Object detections build the basis for all MOT methods of this work. Nevertheless, some approaches (as the method of Section 3.5) operate on short but reliable trajectories, so called *tracklets*, which are fragments of object trajectories. Thus, they assume that detections have already been grouped to tracklets in an initial step.

Given two detections $\mathbf{d}_1, \mathbf{d}_2$, an operation that is frequently used throughout this work is the *Intersection over Union* (IoU) defined as

$$\mathrm{IoU}(\mathbf{d}_1, \mathbf{d}_2) \coloneqq \frac{\mathrm{IA}(\mathbf{d}_1, \mathbf{d}_2)}{\mathrm{UA}(\mathbf{d}_1, \mathbf{d}_2)}\,, \tag{2.38}$$

where $\mathrm{IA}(\mathbf{d}_1, \mathbf{d}_2)$ and $\mathrm{UA}(\mathbf{d}_1, \mathbf{d}_2)$ denote the area of intersection and area of union between the detections $\mathbf{d}_1$ and $\mathbf{d}_2$, respectively.

In the following, we introduce three state-of-the-art object detectors that have been used throughout the experiments[9].

Deformable Part Model (DPM). Among the hand-crafted image descriptors employed to detect or classify objects, the *Histogram of Oriented Gradients* (HOG) [170] has proven to be very powerful. To obtain the descriptor, the image is decomposed

---

[9] An additional, older detector is used only in one experiment (see Section 4.7.10). The interested reader is referred to Dollár *et al.* [169] for details about that detector.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

into connected regions called *cells* (typically of size 8 × 8 pixels). Within each cell, image gradients are computed (for color images, gradients are computed per image channel) and aggregated using a histogram on unsigned gradient directions. Each cell is thus assigned a vector $\mathbf{x} \in \mathbb{R}^9$, representing the histogram using nine equally-sized gradient orientation bins between 0° and 180°. To compensate for local differences in illumination and contrast, neighboring cells are grouped into *blocks* (typically of size 16 × 16 pixels). By concatenating the vectors of the cells that comprise a block, each block B obtains a vector representation $\mathbf{x}_B$. Now each such vector $\mathbf{x}_B$ is normalized (*e.g.*, $l2$-normalized). By concatenating all vectors $\mathbf{x}_B$ for each formed block B, an image descriptor is obtained. By construction, HOG is robust to changes in illumination, shadowing, and small geometric or photometric transformations. HOG descriptors are typically used with a *Support Vector Machine* (SVM) [171] to classify images. A computed SVM weight vector $\mathbf{w}$ can be considered as an image filter, which we call *HOG filter* in the following. When applied to an HOG image, a high response of an HOG filter indicates that an image probably contains an object of the desired class. Yet, for objects that can perform non-rigid deformations, *e.g.*, persons, it is challenging to capture all plausible deformations using HOG descriptors, as they are not invariant to these deformations. To tackle this issue, the *Deformable Part Model* (DPM) [47] takes translations of object parts into account.

In DPM, an object model consists of a root filter together with several object part filters and deformation costs. A root filter is an HOG filter at a coarse resolution that classifies an entire object. It thus corresponds to a simple object detector based on HOG [170]. Object parts are described by additional HOG filters, each using the HOG image in a higher resolution than the root filter. A root filter corresponding to a person and part filters, *e.g.*, for the head, are visualized in Figure 2.10. Each part filter has an expected position relative to the root filter; deviations cause deformation costs. An object is detected by finding the positions where the root and part filters locally maximize a score. A dynamic programming approach using a generalized distance transform is employed to find the optima efficiently. The training phase uses a latent-SVM, *i.e.*, the part locations are considered as latent variables. The approach thus learns the most discriminative filters and the deformation cost model, given the (hidden) object part locations. Finally, a mixture model using multiple object models is applied to an image at multiple scales (using a so-called image pyramid), enabling the detection of objects of different sizes. The image is processed in a sliding-window fashion, checking at all possible positions the existence of an object.

Faster R-CNN (FRCNN). The rise of deep learning has improved object detection substantially. *Faster R-CNN* (FRCNN) [73] is one of the earliest and most effective object detectors based on CNNs.

FRCNN belongs to the class of so-called two-stage object detectors: In the first stage, object proposals are created, which are rough estimates of the image position of an a priori unspecified object. The second stage then verifies these proposals, thereby rejecting false detections and assigning valid detections to the corresponding object class while performing a regression on the localization to obtain more accurate detection boxes (see Figure 2.11).

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

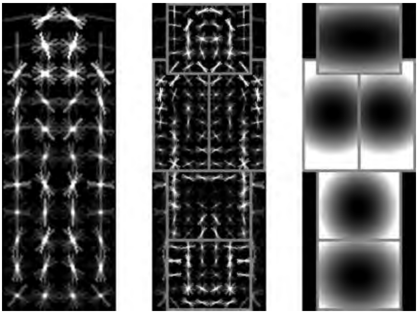


Figure 2.10: Learned filters in order to detect a person [47]. Left: Root filter. Middle: Body part filters. Right: Expected translations of part filters.

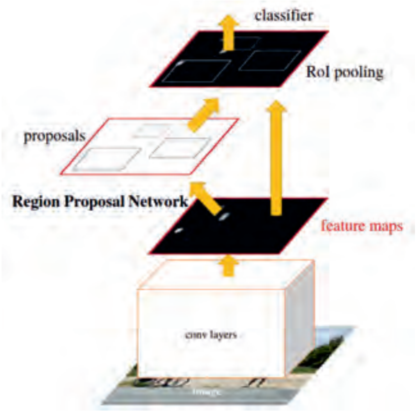


Figure 2.11: The FRCNN architecture [73] consists of two stages: Creating region proposals, followed by a classification and regression of each proposal.

FRCNN creates feature maps using a modification of the neural network VGG-16 [172] (the fully connected layers of VGG-16 are removed). A set of k template bounding boxes of different sizes and aspect ratios are pre-defined, called *anchors*. At each inspected position $\mathbf{p}$ of the feature map, the k anchors are centered at $\mathbf{p}$. Each anchor is assigned a label indicating if there is a corresponding object in the image (positive label) or not (negative label). Given a ground truth box $\mathbf{b}$, the best fitting anchor according to the intersection over union, given by $\mathbf{a}_{\mathbf{b}} := \arg\max_{\mathbf{a}} \mathrm{IoU}(\mathbf{a}, \mathbf{b})$, is assigned a positive label. Moreover, all anchors $\mathbf{a}$ with $\mathrm{IoU}(\mathbf{a}, \mathbf{b}) \geq 0.7$ for some ground truth box $\mathbf{b}$ are assigned the positive label as well. Conversely, an anchor $\mathbf{a}$ with $\mathrm{IoU}(\mathbf{a}, \mathbf{b}) < 0.3$ for all ground truth boxes $\mathbf{b}$ is assigned a negative label. All the other remaining anchors are ignored in the loss function during training. A CNN called *Region Proposal Network* (RPN) is applied to the feature map. It is trained to classify all k anchors at an inspected position $\mathbf{p}$ and to regresses for each anchor the geometry of the corresponding ground

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

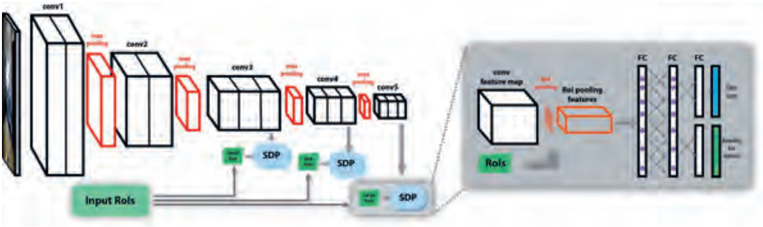


Figure 2.12: SDP architecture [173]. Depending on the object size, the classification of a proposal is performed at a corresponding layer of the neural network.

truth box relative to the anchor. This enables the detection of multiple objects that share common image areas, as each object may have a different anchor that fits best to its shape.

At inference, RPN regresses for each anchor box a corresponding bounding box, called *object proposal*, together with a classification score, called *objectness score*.

In the second phase, for each object proposal with an objectness score above a predefined threshold, a pooling operation called RoI pooling is computed. It discretizes the parts of a feature map of a proposal into a fixed number of equally sized bins and then applies max pooling on each bin. This enables mapping any object proposal to a vector representation of fixed length. Finally, a sub-network performing classification and regression is applied to the pooled object proposal.

FRCNN achieves better accuracies than preceding works as proposal generation (first stage) and verification (second stage) are both neural networks enabling to train both tasks with a shared VGG-16 backbone. Also, FRCNN significantly reduces the required runtime compared to preceding works, as it can handle multiple object scales and sizes without applying the detector on multiple, differently scaled images (so-called image pyramids) or applying multiple filters on the image. Instead, the feature map of RPN is generated once from which the anchor predictions with different scales and sizes are computed.

Scale-Dependent Pooling (SDP). The third object detector *Scale-Dependent Pooling* (SDP) [173] that is used to evaluate the tracking methods of this work is related to FRCNN. Object proposals are derived from existing detection systems with a high recall rate [169, 174]. Again, VGG-16 [172] without the fully-connected layers is applied to the input image, generating feature maps corresponding to the proposals.

Based on their height, object proposals are assigned into one of three possible scale groups. For each scale group s, a layer l_s of VGG-16 is defined (see Figure 2.12). Associated to scale group s is a neural network $\mathrm{SDP}_{\mathrm{head}}(s)$ which performs RoI pooling, object classification, and bounding box regression on the output of layer l_s. Each network $\mathrm{SDP}_{\mathrm{head}}(s)$ is trained and applied only to proposals of corresponding height so that discriminative, scale-dependent features can be obtained. This tackles the problem of objects having different scales and sizes efficiently. As a CNN downscales an image after each pooling layer, small object proposals are processed at an early layer of

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

VGG-16, while for large object proposals, later layers are used as more information is available. Moreover, it allows to speed up the inference time for smaller objects, as not all layers are needed for the computation. Further speedups are achieved by a cascaded rejection classifier using Adaboost [175]. To this end, an object proposal belonging to scale group s is classified at all layers $l_{s'} < l_s$ of VGG-16 using the corresponding feature maps, where s' goes over all smaller scale groups. If a classifier rejects the proposal, the entire computation is stopped since the proposal corresponds to a false detection. It thus addresses the problem of having many proposals.

Finally, Figure 2.13 shows the resulting detections of all three presented detectors when applied to an image frame of the MOT17 dataset [44].

2.7.2 Appearance features

The performance of an MOT method relies to a great extent on the accuracy of the employed pairwise features. Early MOT methods relied mainly on hand-crafted spatio-temporal features [52, 61, 106, 176], while current research has shown that utilizing appearance is key to obtain high-quality MOT results. In this work, we employ appearance-based pairwise features based on two different approaches.

DeepMatching [177] is an algorithm that delivers very accurate, dense pixel correspondences (called *matches*) between two images. The correspondences have many desired properties, *e.g.*, invariance or robustness to global illumination changes, translations, rotations, scalings, and non-rigid deformations. By counting matches between two detection boxes of different frames, DeepMatching poses a robust feature for MOT. The matchings are obtained in an object-category agnostic manner so that the method is stable across poses underrepresented in the training data and partial occlusions.

For the case of people tracking, we also utilize a person re-identification neural network [178]. To this end, each image showing a person is transformed into a discriminative vector representation. This enables an MOT method to decide if two images show the same identity. The network is very robust to non-rigid deformations of persons as well as viewpoint changes.

We briefly introduce both features and refer the interested reader to the literature [177, 178] for further details.

2.7.2.1 ***DeepMatching*** (DM)

DeepMatching[10] [177] delivers correspondences between two images I_1 and I_2, both divided into equally sized squared patches.

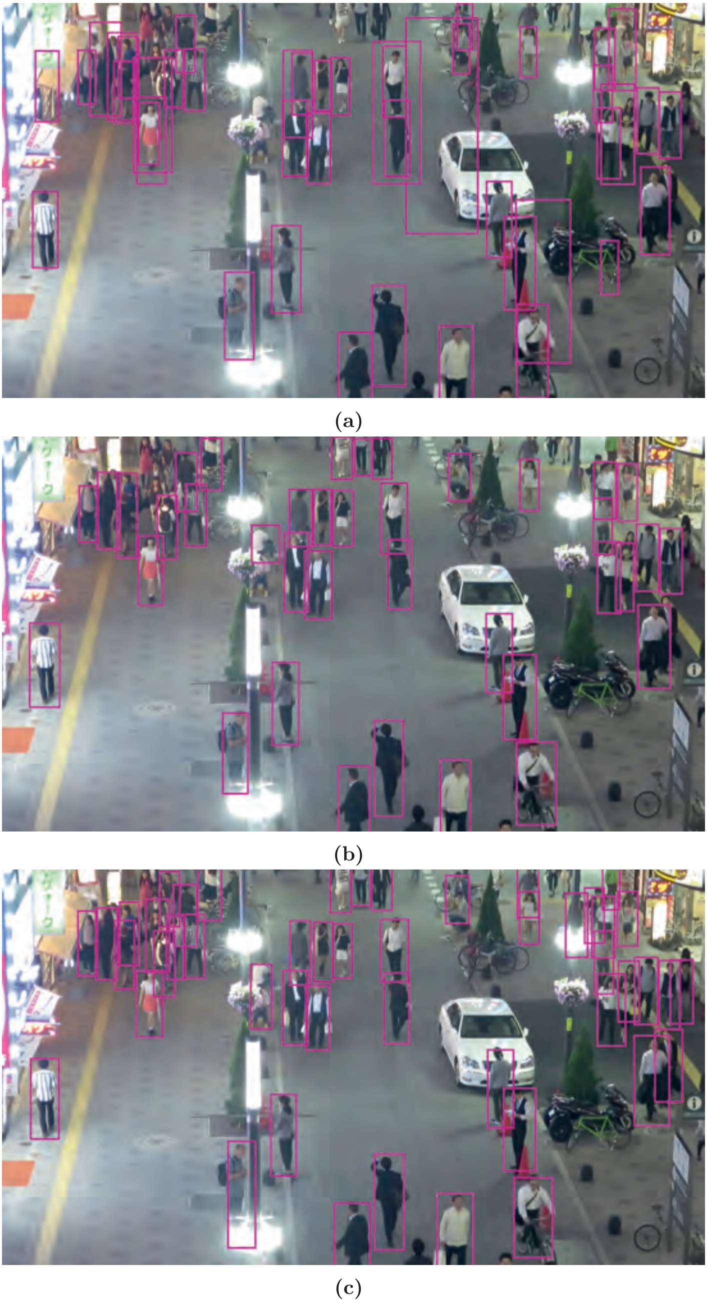
Correspondences are obtained by finding for each patch of I_1 the most similar patch of I_2, while a patch of I_2 can undergo certain transformations to compensate for non-rigid deformations, scalings and rotations (see also Figure 2.14).

DeepMatching thus enables a class-agnostic linkage of detections of two images belonging

---

[10] The prefix "Deep" is used as the algorithm can be described using components of (deep) neural networks. However, DeepMatching has no weights that are learned by a neural network optimizer.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a)

(b)

(c)

Figure 2.13: Detection results for (a) DPM, (b) FRCNN, and (c) SDP. Detections from SDP have the best trade-off between precision and recall.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

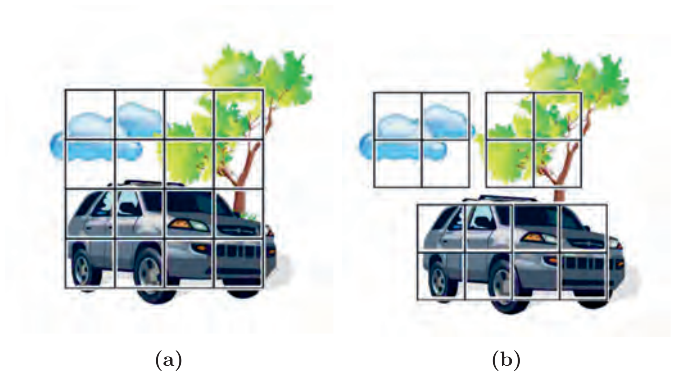


Figure 2.14: DeepMatching computes the correspondences between the patches of an image I_1 (a) to the patches of an image I_2 (b) [177]. Matches are found taking non-rigid transformations into account.

to the same object. Consequently, associations can be performed accurately even in cases of strong occlusions, non-rigid deformations, and between long time distances [63].

The similarity of two patches is determined using an image transformation related to HOG [170], resulting in robustness against noise and global illumination changes. To this end, let I_1 and I_2 be gray-scale images, each of size $w \times h$. For each pixel position $\mathbf{p}$ and $I \in \{I_1, I_2\}$, the image gradient ∇I of I at $\mathbf{p}$ is computed and then non-negatively projected onto the 8 orientations defined by $\mathbf{u}_k \coloneqq (\cos(k\frac{\pi}{4}), \sin(k\frac{\pi}{4}))^\intercal$ for $k \in [8]$. The non-negative projection onto $\mathbb{R}\mathbf{u}_k$ is given by $\tilde{g}_k^I(\mathbf{p}) \coloneqq \operatorname{relu}(\nabla I(\mathbf{p})\mathbf{u}_k)$. Finally, a Gaussian smoothing operation $\mathcal{S}$, a parametrized sigmoid function σ, and a l_2-normalization are applied to $\tilde{\mathbf{g}}^I(\mathbf{p}) = (\tilde{g}_k^I(\mathbf{p}))_{k=1,\dots,8}$ to ensure robustness to noise and varying illumination. Using the mapping $I \mapsto \mathbf{g}^I \coloneqq (l_2 \circ \sigma \circ \mathcal{S})(\tilde{\mathbf{g}}^I)$, I is identified with $\mathbf{g}^I \in \mathbb{R}^{w\times h\times 8}$.

Correspondences between images are then computed hierarchically, using image patches of $\mathbf{g}^{I_1}$ and $\mathbf{g}^{I_2}$. For $d \in \mathbb{N}$, a patch $\mathbf{R} \in \mathbb{R}^{d\times d\times 8}$ of I_1 centered at pixel position $\mathbf{c}_1$ is denoted as $I_{1,d,\mathbf{c}_1} \coloneqq \mathbf{R}$. Likewise, $I_{2,d,\mathbf{c}_2} \in \mathbb{R}^{d\times d\times 8}$ denotes a patch of I_2 centered at $\mathbf{c}_2$.

At the first level of the hierarchical approach, level $l = 0$, the images I_1 and I_2 are decomposed into non-overlapping patches of size 4×4, called *atomic patches*. The similarity of a patch $\mathbf{R}_1 = I_{1,4,\mathbf{c}_1}$ to a patch $\mathbf{R}_2 \in \mathbb{R}^{4\times 4\times 8}$ of I_2, centered at $\mathbf{c}_2$, is then defined as

$$\operatorname{sim}(\mathbf{R}_1, \mathbf{R}_2) \coloneqq R_\lambda \left((I^F_{1,4,\mathbf{c}_1} \star I_2)(\mathbf{c}_2) \right) , \tag{2.39}$$

where R_λ is a non-linear rectification function, $.^F$ denotes the operation of flipping an image horizontally and vertically, and $\star$ denotes the image convolution.

For $d \in \mathbb{N}$, the map $C_{d,\mathbf{c}_1} \coloneqq R_\lambda \circ (I^F_{1,d,\mathbf{c}_1} \star I_2)$ is called *correlation map*. It stores at each position $\mathbf{c}_2$ of I_2 the similarity of a patch $\mathbf{R}_2$ centered at $\mathbf{c}_2$ to $\mathbf{R}_1 = I_{1,d,\mathbf{c}_1}$. Note that $C_{d,\mathbf{c}}$ can be computed by applying a convolutional layer with a pre-defined kernel on I_2, followed by a non-linear activation function.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

However, using atomic patches directly for similarity measurements is error-prone and cannot disambiguate repetitive patterns. To overcome these limitations, a bottom-up procedure is performed to build similarity measurements using larger patches in a hierarchical manner, which shares similarities to the design of convolutional neural networks.

In each hierarchical level $l \geq 1$, the patch dimension parameter d is doubled, resulting in patches $\mathbf{R}^{(1)} \in \mathbb{R}^{d \times d \times 8}$ of I_1 with $d = 2^l \cdot 4$. Thereby, each patch $\mathbf{R}^{(1)}$ is decomposed into 4 equally-sized non-overlapping quadrants, resulting in squared patches of edge-size $\frac{d}{2}$, denoted by $\mathbf{R}_1^{(1)}, \dots, \mathbf{R}_4^{(1)}$ with corresponding centers $\mathbf{c}_1^{(1)}, \dots, \mathbf{c}_4^{(1)}$, respectively. Likewise, a quadratic patch $\mathbf{R}^{(2)}$ of I_2 of edge-length d is decomposed into squared patches $\mathbf{R}_1^{(2)}, \dots, \mathbf{R}_4^{(2)}$ with corresponding centers $\mathbf{c}_1^{(2)}, \dots, \mathbf{c}_4^{(2)}$, respectively. The similarity between patches $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ is then computed using their quadrants. To increase the robustness of the matching, each patch $\mathbf{R}_i^{(2)}$ has a corresponding neighborhood θ_i around its center $\mathbf{c}_i^{(2)}$ which defines admissible translations of $\mathbf{R}_i^{(2)}$. A patch $\mathbf{R}_i^{(2)}$ around center $\mathbf{c}_i^{(2)}$ with translation vector $\boldsymbol{\tau} \in \theta_i$ results in the patch $\mathbf{R}_i^{(2)} \oplus \boldsymbol{\tau} \coloneqq I_{2,\frac{d}{2},\mathbf{c}_i^{(2)}+\boldsymbol{\tau}}$.

Given the features maps of the quadrants $\mathbf{R}_1^{(1)}, \dots, \mathbf{R}_4^{(1)}$ and $\mathbf{R}_1^{(2)}, \dots, \mathbf{R}_4^{(2)}$ which have been computed at the previous level $l-1$, the similarity between $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ is given via

$$\operatorname{sim}(\mathbf{R}^{(1)}, \mathbf{R}^{(2)}) \coloneqq \frac{1}{4} \sum_{i=1}^{4} \max_{\boldsymbol{\tau} \in \theta_i} (I^F_{1,\frac{d}{2},\mathbf{c}_i^{(1)}} \star I_2)(\mathbf{c}_i^{(2)} + \boldsymbol{\tau}) \,. \tag{2.40}$$

The maximization at the right-hand side of Eq. (2.40) finds for each patch $\mathbf{R}_i^{(1)} \coloneqq I_{1,\frac{d}{2},\mathbf{c}_i^{(1)}}$ the optimal translation $\boldsymbol{\tau}$ of $\mathbf{R}_i^{(2)}$ such that the correlation between $\mathbf{R}_i^{(1)}$ and $\mathbf{R}_i^{(2)} \oplus \boldsymbol{\tau}$ is maximized.

By allowing translations via θ_i, the similarity measurement is, up to a certain degree, robust to non-rigid deformations, scalings, and rotations [177].

To ensure efficient computations, a slightly modified version of problem (2.40) is solved in practice. The maximization at each level l is performed using a 3×3 max-pool operation with stride 2. Thus, the correlation map is sub-sampled, allowing the use of a level-independent, constant search domain while effectively increasing the size of the patches that are being correlated. Finally, a non-linear rectification function R_λ is applied. Therefore, the correlation map for a patch $\mathbf{R}^{(1)}$ centered at $\mathbf{c}^{(1)}$ is recursively obtained via

$$C_{d,\mathbf{c}_1} = R_\lambda\Big(\frac{1}{4} \sum_{i=1}^{4} (T_i \circ S \circ P)(C_{\frac{d}{2},\mathbf{c}_i^{(1)}})\Big) \,,$$

where P and S denote the max-pooling and sub-sampling operation, respectively and T_i is a translation operation that is necessary due to the sub-sampling.

After $l_{\max}$ iterations, the correlation maps for all patches up to size $d_{\max} = 2^{l_{\max}} \cdot 4$ have been computed. Starting from the largest patches, the matches between atomic patches can be retrieved using a backtracking algorithm, applied on the correlation maps. Finally, out of these atomic matches, the matches are kept that are local optima with respect to a backtracking score. Thereby, the score between two atomic patches is defined as the sum of all patch similarities along the backtracking path from the atomic patches up to the patches of size $d_{\max} \times d_{\max}$. Note that a correspondence between two

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

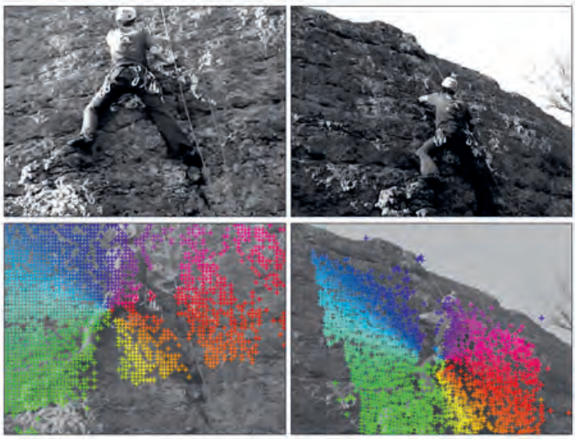


Figure 2.15: A qualitative result of DeepMatching [177]. Despite geometric transformations, DeepMatching delivers accurate dense correspondences.

patches can appear within different backtracking paths, resulting in different scores. DeepMatching thus employs a filtering procedure, keeping only the correspondences with the highest score. A qualitative result of DeepMatching is shown in Figure 2.15. The feature is used in the proposed MOT methods of Chapter 3 and Chapter 4.

2.7.2.2 Person Re-Identification

The accuracy of an MPT method depends to a great extent on the ability to automatically assess whether two detection boxes show the same person. Thus, this problem is strongly related to *person re-identification*, an ongoing research topic of computer vision. It has seen tremendous progress by utilizing deep neural networks [179].

Person re-identification is often applied in a setup where either the two images are taken from different cameras or at different epochs. The temporal distance is usually in the range of seconds to minutes. Accordingly, computed similarities need to be robust against occurring intra-class variations due to changing human poses and background, varying camera angles, and different kinds of noise. A person re-identification method capable of dealing with these challenges is thus well suited as a basis for pairwise features in MPT. Note that the similarities are computed without using any spatial or temporal context. Instead, it is based solely on visual information.

The re-identification problem is commonly tackled by finding a similarity measure on a suitable embedding space that reflects semantic similarity.

To this end, a neural network $h_{\mathbf{w},\mathbf{b}} : \mathbb{R}^i \rightarrow \mathbb{R}^f$ needs to be determined that maps an i-dimensional image to an f-dimensional feature space such that any two images $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^i$ showing the same person have embedding vectors $h_{\mathbf{w},\mathbf{b}}(\mathbf{x}_1), h_{\mathbf{w},\mathbf{b}}(\mathbf{x}_2)$ close in the embedding space, measured in terms of a similarity function $\mathrm{sim} : \mathbb{R}^f \times \mathbb{R}^f \rightarrow \mathbb{R}$.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Fixing a neural network architecture in terms of a graph $\mathcal{G}$ (see Definition 2.17), the goal is to determine the best weights and biases $(\mathbf{w}, \mathbf{b})$, resulting in the map $h_{\mathbf{w},\mathbf{b}} \in \mathcal{H}(\mathcal{G})$.

Baseline. Chapter 4 uses a re-identification method [178] that builds upon a standard procedure by training an identity classifier on a finite dataset $T \subset \mathbb{R}^i$ of labeled images showing n_P persons, where each sample $\mathbf{x} \in T$ has a corresponding label $y \in [n_P]$.

For a neural network $h_{\mathbf{w},\mathbf{b}} : \mathbb{R}^i \to \mathbb{R}^f$, we consider a matrix $\mathbf{W}' \in \mathbb{R}^{n_P \times f}$ and vector $\mathbf{b}' \in \mathbb{R}^{n_P}$. The probability of a label $l \in [n_P]$ for a given image $\mathbf{x}$ is computed as

$$p_{\mathbf{W}',\mathbf{b}',h_{\mathbf{w},\mathbf{b}}}(\hat{y} = l \mid \mathbf{x}) = \frac{\exp\left(\mathbf{W}'_{[l,:]} h_{\mathbf{w},\mathbf{b}}(\mathbf{x}) + b'_l\right)}{\sum_{\tilde{l}=1}^{n_P} \exp\left(\mathbf{W}'_{[\tilde{l},:]} h_{\mathbf{w},\mathbf{b}}(\mathbf{x}) + b'_{\tilde{l}}\right)} . \tag{2.41}$$

The multi-class classification task is then to predict for each given image $\mathbf{x}$ the correct class. Optimal parameters $\mathbf{W}'$, $\mathbf{b}'$, $\mathbf{w}$ and $\mathbf{b}$ are obtained using cross-entropy loss. Finally, the embedding vectors are l_2-normalized, resulting in a neural network $\underline{h}_{\mathbf{w},\mathbf{b}}$. To simplify notation, we denote the trained network $\underline{h}_{\mathbf{w},\mathbf{b}}$ as h^{B}. The similarity measure between embedding vectors is then given as

$$\operatorname{sim}(\mathbf{x}_1, \mathbf{x}_2) \coloneqq \langle h^{\mathrm{B}}(\mathbf{x}_1), h^{\mathrm{B}}(\mathbf{x}_2) \rangle . \tag{2.42}$$

The embedding vectors h^{B} use the ResNet-50 CNN [135], while the last layer is replaced by a fully connected layer of dimension $f = 512$, followed by a batch normalization.

DG-Net. In order to improve robustness against intra-class variation, the dataset T can be augmented by synthetic images during the training phase. The current state of the art to create synthetic images such that its distribution is similar to the distribution of the images of the dataset is to employ a generative adversarial network [180]. However, assigning identity labels to synthetic images is often ambiguous and might thus deteriorate the quality of the similarity function. Instead, the employed re-identification system [178] called DG-Net jointly learns the discriminative task and the creation of synthetic images with high intra-class variation while preserving the identities of the dataset. The labeling issues are tackled by assigning synthetic images soft labels.

A brief sketch of the coupled image generation and discriminative learning part is presented in the following. The interested reader is referred to the work of Zhedong *et al.* [178] for further details.

The system involves several neural networks that are jointly trained: an *appearance encoder* $E_{\mathrm{app}} : \mathbb{R}^i \to \mathbb{R}^a$, a *structure encoder* $E_{\mathrm{str}} : \mathbb{R}^i \to \mathbb{R}^s$, a *decoder* $G : \mathbb{R}^a \times \mathbb{R}^s \to \mathbb{R}^i$, and a discriminator $D : \mathbb{R}^i \to [0, 1]$.

Conceptually, the idea is to disentangle an image of a person into an appearance code (using E_{app}) describing identity-related cues such as clothing colors and a structure code (using E_{str}) describing geometric information such as positions and pose as well as hair and body sizes (see Figure 2.16).

In order to focus on the structural information, the input image is converted to grayscale when passed through E_{str}. The disentanglement enables the generation of images

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

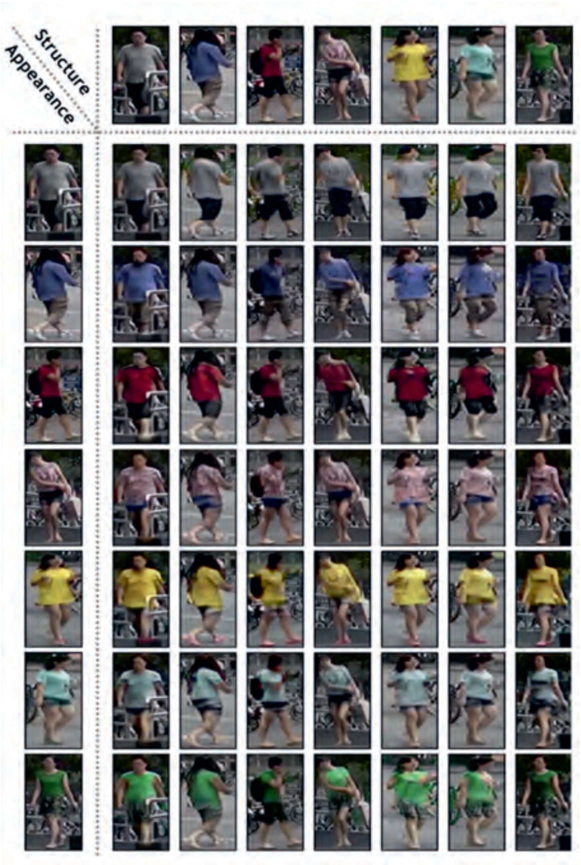


Figure 2.16: Using a disentanglement of a person image into appearance code and structure code enables to augment the training data by combining these codes from different person images [178].

(using G) with much higher intra-class variation, *e.g.*, by creating images of the same person but with a varying pose. Accordingly, the trained system becomes robust to pose variations, changing lighting conditions and more precise to the re-identification task.

In order to couple image generation and the creation of embedding vectors that can cope with intra-class variations, an embedding map $h^{\mathrm{DG}} : \mathbb{R}^i \to \mathbb{R}^f$ is used. h^{DG} is a neural network that shares its ResNet-50 backbone with the appearance encoder E_{app}.

Several losses are employed to ensure correct disentanglement and identity preservation. In the following, let $\mathbf{x}^{\mathfrak{a}} := E_{\mathrm{app}}(\mathbf{x})$ and $\mathbf{x}^{\mathfrak{s}} := E_{\mathrm{str}}(\mathbf{x})$ denote the appearance code and structure code for image $\mathbf{x} \in \mathbb{R}^i$, respectively. We denote by $T \subset \mathbb{R}^i$ a training set. For

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

each $\mathbf{x} \in T$, we denote by $y(\mathbf{x}) \in [n_P]$ the corresponding label. The appearance code and structure code of an image $\mathbf{x} \in \mathbb{R}^i$ must contain all information to reconstruct $\mathbf{x}$, which is enforced by the loss

$$L_0 \coloneqq \frac{1}{|T|} \sum_{\mathbf{x} \in T} \left\| \mathbf{x} - G(\mathbf{x}^{\mathfrak{a}}, \mathbf{x}^{\mathfrak{s}}) \right\|_1 . \tag{2.43}$$

The decoder should reconstruct an image $\mathbf{x}$ of a person $y \in [n_P]$ if the structure code $\mathbf{x}^{\mathfrak{s}}$ and the appearance code $\mathbf{x}_2^{\mathfrak{a}}$ are used, where $\mathbf{x}_2$ is another image of person y. To this end, we define the set of image pairs showing the same person $T_{\text{same}} \coloneqq \{(\mathbf{x}_1, \mathbf{x}_2) \in T^2 \mid y(\mathbf{x}_1) = y(\mathbf{x}_2)\}$. The reconstruction is then ensured using the following loss:

$$L_1 \coloneqq \frac{1}{|T_{\text{same}}|} \sum_{(\mathbf{x}_1, \mathbf{x}_2) \in T_{\text{same}}} \left\| \mathbf{x}_1 - G(\mathbf{x}_2^{\mathfrak{a}}, \mathbf{x}_1^{\mathfrak{s}}) \right\|_1 . \tag{2.44}$$

Given an image $\mathbf{x}'$ decoded from structure code $\mathbf{x}_1^s$ and appearance code $\mathbf{x}_2^a$, computing the disentangled encodings of $\mathbf{x}'$ should recover the same codes, which is ensured by

$$L_2 \coloneqq \frac{1}{|T|^2} \sum_{(\mathbf{x}_1, \mathbf{x}_2) \in T^2} \left\| \mathbf{x}_2^{\mathfrak{a}} - E_{\text{app}}(G(\mathbf{x}_2^{\mathfrak{a}}, \mathbf{x}_1^{\mathfrak{s}})) \right\|_1 \tag{2.45}$$

and

$$L_3 \coloneqq \frac{1}{|T|^2} \sum_{(\mathbf{x}_1, \mathbf{x}_2) \in T^2} \left\| \mathbf{x}_1^{\mathfrak{s}} - E_{\text{str}}(G(\mathbf{x}_2^{\mathfrak{a}}, \mathbf{x}_1^{\mathfrak{s}})) \right\|_1 . \tag{2.46}$$

To obtain synthetic images looking similar to the images of the dataset, an adversarial loss [180] is used that is maximized over D and minimized over G.

$$L_4 \coloneqq \frac{1}{|T|} \sum_{\mathbf{x} \in T} \log(D(\mathbf{x})) + \frac{1}{|T|^2} \sum_{(\mathbf{x}_1, \mathbf{x}_2) \in T^2} \log(1 - D(G(\mathbf{x}_1^{\mathfrak{a}}, \mathbf{x}_2^{\mathfrak{s}}))). \tag{2.47}$$

By swapping structure and appearance code between two images, $\mathcal{O}(|T|^2)$ many different synthetic images can be created. Still, it is not straightforward to assign a discrete label to synthetic images. Therefore, the baseline re-identification system h^{B} is used to construct soft labels. The aim is then to train a neural network that mimics the similarity function of Eq. (2.42). To this end, the KL-divergence between the posterior class probabilities (see Eq. (2.41)) given by h^{B} and h^{DG} is minimized.

To simplify notation, we omit in the following weights and biases and denote for embedding h by $p_h(c|\mathbf{x})$ the probability $p_{\mathbf{W},\mathbf{b},h}(\hat{y} = l|\mathbf{x})$ defined in Eq. (2.41). Additionally, we define $\mathbf{x}_{1:\mathfrak{a}}^{2:\mathfrak{s}} \coloneqq G(\mathbf{x}_1^{\mathfrak{a}}, \mathbf{x}_2^{\mathfrak{s}})$. The next loss then ensures that h^{DG} mimics h^{B}:

$$L_5 \coloneqq \frac{-1}{n_P |T|^2} \sum_{\substack{(\mathbf{x}_1, \mathbf{x}_2) \in T^2 \\ l \in [n_P]}} p_{h^{\text{B}}}(l \mid \mathbf{x}_{1:\mathfrak{a}}^{2:\mathfrak{s}}) \log \left(\frac{p_{h^{\text{DG}}}(l \mid \mathbf{x}_{1:\mathfrak{a}}^{2:\mathfrak{s}})}{p_{h^{\text{B}}}(l \mid \mathbf{x}_{1:\mathfrak{a}}^{2:\mathfrak{s}})} \right) . \tag{2.48}$$

Thereby, only the weights of h^{DG} are allowed to be changed, while h^{B} is kept fixed.

In order to improve the discriminative accuracy of the network h^{DG}, three additional losses are added. For images of the training set, the network should output the correct identities, thus:

$$L_6 \coloneqq \frac{1}{|T|} \sum_{\mathbf{x} \in T} - \log(p_{h^{\text{DG}}}(y(\mathbf{x}) \mid \mathbf{x})) . \tag{2.49}$$

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

If the structure code and appearance code are mixed, the appearance should determine the identity:

$$L_7 := \sum_{(\mathbf{x}_1,\mathbf{x}_2)\in T^2} -\log(p_{h^{\mathrm{DG}}}(y(\mathbf{x}_1) \mid \mathbf{x}^{2:s}_{1:a}))\,. \tag{2.50}$$

Another loss is added with a small weight that helps the network to identify a person if the appearance is changed but the structure is kept. This helps to incorporate fine-graded features such as a person's hair or body size into embedding vectors.

$$L_8 := \sum_{(\mathbf{x}_1,\mathbf{x}_2)\in T^2} -\log(p_{h^{\mathrm{DG}}}(y(\mathbf{x}_2) \mid \mathbf{x}^{2:s}_{1:a}))\,. \tag{2.51}$$

The final loss used to train[11] the neural networks is then given as

$$L = \sum_{i=0}^{8} \lambda_i L_i\,, \tag{2.52}$$

for hyperparameters $\lambda_0, \dots, \lambda_8 \in \mathbb{R}$. The resulting embedding vector h^{DG} is then used as in Eq. (2.42) to calculate the similarity function. We refer to Zheng *et al.* [178] for further details about the training procedure and the architecture of the employed neural networks. The re-identification system is used in the tracking method proposed in Chapter 4.

Person re-identification metrics. Finally, we present two state-of-the-art metrics [181] to evaluate the performance of a re-identification method. Given a detection $\mathbf{d}$, a re-identification method outputs a distance to any other detection $\mathbf{d}'$ according to the learned embedding vectors.

Hence given a query image of a person defined by a detection $\mathbf{d}$ and a gallery set G of collected images, each defined by a detection, we obtain the *best matching* image via $\mathbf{d}^* = \arg\min_{\mathbf{d}'\in G} f(\mathbf{d}, \mathbf{d}')$, where f is a suitable distance metric on the corresponding embedding vectors. Now the relative number of correct best matchings for a query set Q denotes the Rank-1 metric.

For the second metric, we assume that $G = \{\mathbf{d}_1, \dots, \mathbf{d}_n\}$ is ordered with increasing distance to detection $\mathbf{d}$. Let l be a label function such that for each $i \in [n]$, $l(\mathbf{d}, \mathbf{d}_i) = 1$ if $\mathbf{d}$ and $\mathbf{d}_i$ depict the same person, and 0 otherwise. The *precision at k* given by

$$P@k(\mathbf{d}) := \frac{1}{k}\sum_{i=1}^{k} l(\mathbf{d}, \mathbf{d}_i) \tag{2.53}$$

computes for the k best matchings the relative number of correct matchings. The expected precision, called *average precision* is given as

$$AveP(\mathbf{d}) := \sum_{k=1}^{n} \frac{P@k(\mathbf{d})l(\mathbf{d}, \mathbf{d}_k)}{a_{\mathbf{d}}}\,, \tag{2.54}$$

where $a_{\mathbf{d}}$ counts the valid matches, *i.e.*, the number of detections $\mathbf{d}'$ with $l(\mathbf{d}, \mathbf{d}') = 1$.

Finally, the *mean Average Precision* (mAP) metric is obtained via

$$\mathrm{mAP} := \frac{1}{|Q|}\sum_{\mathbf{d}\in Q} AveP(\mathbf{d})\,. \tag{2.55}$$

---

[11] To simplify notation, we omitted the dependence of loss L on the weights that are optimized during training for $h^{\mathrm{DG}}, E_a, E_s, G$ and D.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 2.1: Characteristics of the datasets used in this work. Density is measured in average number of pedestrians per frame. The MOT15 dataset has on average fewer frames per sequence and fewer pedestrians per frame, compared to MOT16/MOT17. Note that MOT17 uses the same sequences as MOT16 but three different detectors.

Dataset	Mean length [frames]	Density	Detector
MOT15	512.9	9	ACF
MOT16	802.5	26.1	DPM
MOT17	802.5	26.1	DPM/FRCNN/SDP

2.7.3 Datasets

To assess the quality of an MOT method, several datasets have been published [2, 43, 44, 182–184]. However, for some of the older datasets, detections were not provided so that researchers created their own set of detections, making it difficult to tell apart the improvement coming from the detector and the MOT method. Also, some works assess their MOT methods on self-created ground truth labels, hindering meaningful insights from quantitative metrics, as it is crucial to create ground truth labels using the same protocol. The impact of these issues has been analyzed and quantified by Milan *et al.* [185].

Several benchmarks have been created [2, 43, 44] to tackle these issues. They ensure a standardized evaluation, as they provide detections for all sequences that competing MOT methods have to take as input. Ground-truth labels are publicly available only for the training set. The evaluation on the test set is calculated on a public server. This ensures that only a few evaluations per tracking method can be performed so that parameters of an MOT method cannot be optimized on the test set.

One of the most recent datasets following this principle are the datasets MOT15 [43], MOT16 [44], and MOT17 [44]. They are used in this work to assess the performance of the proposed methods. The datasets comprise mainly outdoor recordings of persons walking in natural street scenes. In more detail, MOT15, MOT16, and MOT17 comprise 11, 7, and 7 sequences in the training set as well as in the test, respectively. MOT15 uses the *Aggregate Channel Features* (ACF) detector [169], while MOT16 employs DPM [170] detections. MOT17 shares the same set of sequences as MOT16, but the detectors DPM, FRCNN, and SDP are applied for each sequence, resulting in three different inputs for a tracking method per sequence. This setup makes it possible to analyze the influence of the input quality on the tracking result.

The recordings of the datasets differ in the frame rate, from 7 up to 30 *frames per second* (fps). The resolution ranges from 640 × 480 up to Full HD. Sequences are recorded by a static or moving camera (the camera has been placed either on a stroller, mounted on a moving car, or held by a person). Table 2.1 shows that sequences of MOT16 and MOT17 are more crowded and contain on average more image frames. Further analyses of the datasets are provided by Milan *et al.* [44] and Leal-Taixe *et al.* [43]. Exemplary images of the datasets are shown in Figure 2.17.

To ensure meaningful comparisons, parameters of a tracking method are optimized

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

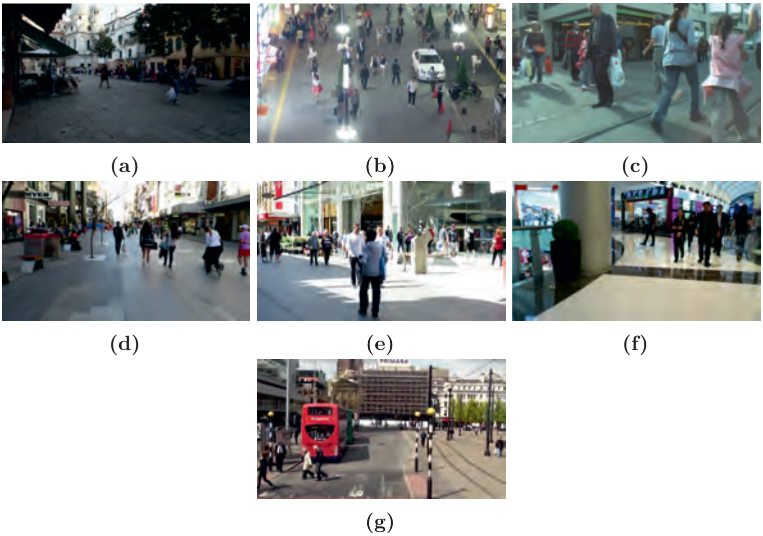


Figure 2.17: Scenes from the seven test sequences of the MOT16 and MOT17 datasets. Only the sequences shown in (a), (b), and (e) are recorded by a static camera. The sequences differ a lot in complexity. The sequence of (b) depicts around 70 persons per frame. The distance of people to the camera varies from close-by (c) & (e), to far-away (g), making it difficult to define a stable distance measurement as only 2D information is provided. Due to the different camera angles, also the degree of occlusion is very different for the sequences. Note that also the lighting differs a lot between the different sequences.

using the entire training data and held fixed when applied to a sequence of the test set.

2.7.4 MOT metrics

In order to assess the quality of the trajectories created by an MOT method, several metrics[12] have been proposed. Following Ristani *et al.* [186], they are categorized into two different types of metrics:

- **Event-based metrics**: They are based on the frequency of certain error types produced by an MOT method, which is useful to analyze and detect weaknesses.
- **Identity-based metrics**: They assess how well each ground truth trajectory is described by a corresponding computed trajectory. This is particularly important for applications in which each trajectory must follow exactly one identity, *e.g.*, when analyzing the performance of individual athletes.

---

[12] The terminology *metric* is commonly used in the context of computer vision and tracking to denote a function mapping to $\mathbb{R}$.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

The essential difference between the two categories is how computed trajectories are assigned to ground truth trajectories.

Event-based metrics. In the following, we introduce the definitions of event-based metrics termed CLEAR MOT [187] and its extensions [188], based on the implementation provided by the datasets MOT15 [43] and MOT16/MOT17 [44].

Let $\Gamma = \{\gamma_1, \dots, \gamma_n\}$ be the tracking result of an MOT method applied to a sequence and let $D_\mathfrak{f}$ denote the corresponding set of detections at frame $\mathfrak{f} \in [n_\mathrm{R}]$ so that input detections not contained in any trajectory are discarded. Additionally, let $D^\star_\mathfrak{f}$ be the set of ground truth detections at frame $\mathfrak{f} \in [n_\mathrm{R}]$. A pairing of detections $(\mathbf{d}_1, \mathbf{d}_2)$ is defined to be *compatible* if $\mathrm{IoU}(\mathbf{d}_1, \mathbf{d}_2) \geq 0.5$.

Event-based metrics utilize a frame-wise correspondence (matching) between computed and ground truth trajectories, using a bipartite graph $\mathcal{G}_\mathfrak{f} = (D_\mathfrak{f} \sqcup D^\star_\mathfrak{f}, \mathcal{E}_\mathfrak{f}, w_\mathfrak{f})$ for each frame $\mathfrak{f} \in [n_\mathrm{R}]$. Thereby, $\mathcal{E}_\mathfrak{f} := \{\mathbf{d}_\mathrm{det}\mathbf{d}_\mathrm{gt} \mid \mathbf{d}_\mathrm{det} \in D_\mathfrak{f}, \mathbf{d}_\mathrm{gt} \in D^\star_\mathfrak{f}, (\mathbf{d}_\mathrm{det}, \mathbf{d}_\mathrm{gt}) \text{ compatible}\}$. The weight of an edge $e = \mathbf{d}_\mathrm{det}\mathbf{d}_\mathrm{gt} \in \mathcal{E}_\mathfrak{f}$ is defined as $w_\mathfrak{f}(e) = \mathrm{IoU}(\mathbf{d}_\mathrm{det}, \mathbf{d}_\mathrm{gt})$.

For each frame $\mathfrak{f} \in [n_\mathrm{R}]$, a matching $M_\mathfrak{f}$ is constructed. Initially, $M_\mathfrak{f} = \emptyset$. Each edge $e = \mathbf{d}_\mathrm{det}\mathbf{d}_\mathrm{gt} \in \mathcal{E}_\mathfrak{f}$ connects $\mathbf{d}_\mathrm{det}$ of a computed trajectory γ with $\mathbf{d}_\mathrm{gt}$ of a ground truth trajectory $\gamma^\star$. If γ has been linked to $\gamma^\star$ in frame $\mathfrak{f} - 1$, this correspondence is maintained in frame $\mathfrak{f}$. Thus if $\gamma(\mathfrak{f}-1)\gamma^\star(\mathfrak{f}-1) \in M_{\mathfrak{f}-1}$, the matching set is updated to $M_\mathfrak{f} := M_\mathfrak{f} \cup \{e\}$. This is done even if there is a better fitting matching to $\mathbf{d}_\mathrm{gt}$, as this procedure ensures temporal consistency within the evaluation. For the remaining set of edges $\mathcal{E}' = \mathcal{E}_\mathfrak{f} \setminus M_\mathfrak{f}$, an optimal matching is computed on $\mathcal{G}_\mathfrak{f}[\mathcal{E}']$ using the Hungarian algorithm, resulting in the final matching set $M_\mathfrak{f}$. The matching procedure is illustrated in Figure 2.18.

Definition 2.52. For a sequence with n_R frames, we define several metrics with respect to the correspondences $M = \cup_{\mathfrak{f}=1}^{n_\mathrm{R}} M_\mathfrak{f}$.

- *False Positives* (FP) denotes the number of detections of computed trajectories that have not been matched to a ground truth trajectory. The lower the score, the better.
- *False Negatives* (FN) denotes the number of ground truth detections that do not have a corresponding detection from a computed trajectory. The lower the score, the better.
- Consider a matched detection of a computed trajectory γ at frame $\mathfrak{f}$. Let $\triangle\mathfrak{f} \in \mathbb{N}$ be minimal such that $\gamma(\mathfrak{f} - \triangle\mathfrak{f})$ is matched, too. If the ground truth trajectory matched to $\gamma(\mathfrak{f})$ differs from the ground truth trajectory matched to $\gamma(\mathfrak{f} - \triangle\mathfrak{f})$, the error is counted as an *Identity* (ID) *switch*. The number of ID switches for the entire sequence is denoted as *ID Switches* (IDS). The lower the score, the better.
- The metrics are combined to the *Multiple Object Tracking Accuracy* (MOTA):

  $$\mathrm{MOTA} = 1 - \frac{\mathrm{FP} + \mathrm{FN} + \mathrm{IDS}}{|D^*|}\,.$$

  Thus, MOTA is a value in the interval $(-\infty, 1]$. The maximal value indicates the optimal MOT result.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

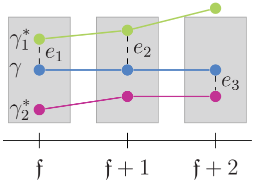


Figure 2.18: Illustration of the matching procedure for the event-based metrics with a computed trajectory γ (blue) and two ground truth trajectories γ_1^* and γ_2^*. All nodes within the gray area at a respective frame are compatible with the computed trajectory. The green and blue trajectories are matched at frame $\mathfrak{f}$. Consequently, they are also matched in frame $\mathfrak{f}+1$, even though the purple trajectory is much closer. In frame $\mathfrak{f}+2$, the detection of the green trajectory is not compatible. Hence, the detection of the purple trajectory is matched to the blue trajectory. The resulting matching $M = \{e_1, e_2, e_3\}$ implies one ID switch and three false negatives.

- The metrics *Mostly Tracked* (MT), *Partially Tracked* (PT), and *Mostly Lost* (ML) assess the relative coverage of ground truth trajectories by computed trajectories. That is, MT, PT, and ML count the relative number of ground truth trajectories that have corresponding matches for more than 80%, between 20% and 80%, and less than 20% of its length, respectively. Accordingly, the aim is to design an MOT method with a low ML and PT value and a high MT value.

Identity-based metrics. Identity-based metrics differ essentially from event-based metrics by establishing a correspondence between entire trajectories instead of matching detections to ground truth on a frame-by-frame basis. Conceptually, the IDF1 score [186] assesses the reliability of tracking a particular object and is evaluated for all objects of a sequence.

To this end, a bipartite graph $(\mathcal{V}_C \sqcup \mathcal{V}_{GT}, \mathcal{E})$ is constructed such that $\mathcal{V}_C$ contains one node for each computed trajectory and for each ground truth trajectory γ^* an additional node $f_{\gamma^*}^-$ that represents false negative trajectories. Likewise, $\mathcal{V}_{GT}$ contains one node for each ground truth trajectory and for each computed trajectory γ an additional node f_γ^+ that represents false positive trajectories. We denote the set of all additional nodes f_γ^+ and $f_{\gamma^*}^-$ by $\mathcal{V}_{FP}$ and $\mathcal{V}_{FN}$, respectively. Trajectories $\gamma \in \mathcal{V}_C \setminus \mathcal{V}_{FP}$ and $\gamma^* \in \mathcal{V}_{GT} \setminus \mathcal{V}_{FN}$ are connected by an edge $e = (\gamma, \gamma^*)$ if γ and γ^* have some temporal overlap. Each trajectory $\gamma \in \mathcal{V}_C \setminus \mathcal{V}_{FP}$ is connected with f_γ^+. Likewise, $\gamma^* \in \mathcal{V}_{GT} \setminus \mathcal{V}_{FN}$ is connected with $f_{\gamma^*}^-$.

For a sequence of length n_R, the set of frames for which a trajectory γ contains detections, denoted as $\operatorname{supp}(\gamma) := \{\mathfrak{f} \in [n_\mathrm{R}] \mid \gamma(\mathfrak{f}) \neq \emptyset\}$, is the *support* of γ. For each computed trajectory γ and ground truth trajectory γ^*, we define

$$\operatorname{supp}(f_\gamma^+) := \emptyset \quad \text{and} \quad \operatorname{supp}(f_{\gamma^*}^-) := \emptyset. \tag{2.56}$$

Now given an edge $e = (\gamma, \gamma^*) \in \mathcal{E}$ with $\gamma \in \mathcal{V}_C$ and $\gamma^* \in \mathcal{V}_{GT}$ and assume that $\gamma \notin \mathcal{V}_{FP} \vee \gamma^* \notin \mathcal{V}_{FN}$ holds. Then for frame $\mathfrak{f} \in [n_\mathrm{R}]$, we compute if detection $\gamma^*(\mathfrak{f})$ and

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

detection $\gamma(\mathfrak{f})$ cannot be matched:

$$m(\gamma, \gamma^*, \mathfrak{f}) \coloneqq \begin{cases} 0, & (\mathfrak{f} \in \operatorname{supp}(\gamma) \cap \operatorname{supp}(\gamma^*)) \text{ and } ((\gamma(\mathfrak{f}), \gamma^*(\mathfrak{f})) \text{ compatible}), \\ 1, & \text{otherwise.} \end{cases} \tag{2.57}$$

For $\gamma \in \mathcal{V}_{FP} \land \gamma^* \in \mathcal{V}_{FN}$, we set $m(\gamma, \gamma^*, \mathfrak{f}) \coloneqq \infty$.

Thus given an edge $e = (\gamma, \gamma^*) \in \mathcal{E}$, we obtain the number of false positives, denoted as $\mathrm{IDFP}(\gamma)$ and false negatives, denoted as $\mathrm{IDFN}(\gamma^*)$ using the match e, *i.e.*,

$$\mathrm{IDFP}(e) \coloneqq \sum_{\mathfrak{f} \in \operatorname{supp}(\gamma)} m(\gamma, \gamma^*, \mathfrak{f})\,, \quad \mathrm{IDFN}(e) \coloneqq \sum_{\mathfrak{f} \in \operatorname{supp}(\gamma^*)} m(\gamma, \gamma^*, \mathfrak{f})\,. \tag{2.58}$$

We define for each potential match $e = (\gamma, \gamma^*) \in \mathcal{E}$ the weight $w(e) \coloneqq \mathrm{IDFP}(\gamma) + \mathrm{IDFN}(\gamma^*)$. A minimum cost optimal matching M then provides the basis for the assessment of the tracking result. To this end, we compute the number of false negative ID matches (IDFN) and false positive ID matches (IDFP) according to M as

$$\mathrm{IDFP} \coloneqq \sum_{e \in M} \mathrm{IDFP(e)}, \quad \mathrm{IDFN} \coloneqq \sum_{e \in M} \mathrm{IDFN(e)}\,. \tag{2.59}$$

In addition, we obtain the number true positive ID matches (IDTP) via

$$\mathrm{IDTP} \coloneqq \sum_{(\gamma,\gamma^*) \in M} |\operatorname{supp}(\gamma)| - \mathrm{IDFP} = \sum_{(\gamma,\gamma^*) \in M} |\operatorname{supp}(\gamma^*)| - \mathrm{IDFN}\,. \tag{2.60}$$

Finally, we establish the *ID Precision* (IDP) metric, *ID Recall* (IDR) metric, and the *ID F1* (IDF1) metric:

$$\mathrm{IDP} \coloneqq \frac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFP}}, \; \mathrm{IDR} \coloneqq \frac{\mathrm{IDTP}}{\mathrm{IDTP} + \mathrm{IDFN}}, \; \mathrm{IDF1} \coloneqq \frac{2\mathrm{IDTP}}{2\mathrm{IDTP} + \mathrm{IDFN} + \mathrm{IDFP}}\,. \tag{2.61}$$

The metrics can be interpreted as follows:

- IDP represents the relative number of detections from computed trajectories that are matched to detections from ground truth trajectories (according to M).
- IDR represents the relative number of ground truth detections that are matched to detections of computed trajectories (according to M).
- The IDF1 metric is the F1 score between IDP and IDR.

The higher the IDP, IDR, and IDF1 metrics, the better. We refer to Ristani *et al.* for further details [186].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

3 HO-MOT with Signal Fusion[1]

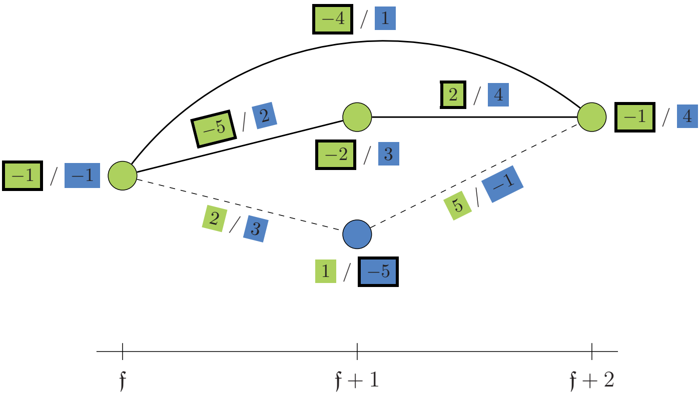


Figure 3.1: Illustration of connections between input signals in three consecutive video frames together with an optimal labeling according to a weighted graph labeling problem. Each node is assigned a label, indicated by the colored nodes. Node and edge weights between equally labeled nodes are summed up using label-dependent weights (indicated by colored node and edge values) to evaluate an assignment. This allows fusing multiple signals, *e.g.*, signals that stem from different sensors. Solid edges represent active connections, dashed edges represent inactive connections between the input signals. All weights framed in black contribute to the objective value of the depicted graph labeling.

This chapter proposes a method for HO-MOT using signal fusion. The performance of tracking-by-detection approaches heavily relies on the quality of the computed object detections. However, detectors are still error-prone (see Section 1.3.1) and pairwise costs between object detections are often misleading in challenging situations (see Section 1.3.2). Correctly interpreting and compensating these errors in the data association part is difficult.

By reducing the dependence on input detections, this chapter addresses the drawback of the tracking-by-detection paradigm. To this end, a fusion approach for implicit

---

[1] This chapter contains text, images, and results of previously published work [55, 121, 122].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

HO-MOT is proposed based on a weighted graph labeling problem. It enables to augment object detections with additional complementary signals. Solving the graph labeling problem delivers globally optimal input-to-identity assignments across all signals. The underlying optimization problem is an $\mathcal{NP}$-hard BQP. An approximate solver adapted to the problem formulation is proposed that delivers solutions close to the optimum in practice.

The advantage of the fusion formulation is demonstrated in two settings: (i) By combining people detections with head detections, partially occluded persons can be tracked more reliably, as in such a situation a person's head is often still visible. Also, trajectories that do not have assigned head detections are likely to collect false positive detections. Consequently, the fusion helps to increase tracking precision and recall. Compared to the accuracy when using only person detections as input, the fusion approach significantly improves tracking results. (ii) The second setup demonstrates that the fusion formulation can also be applied across different modalities by using video data and an *Inertial Measurement Unit* (IMU) attached to the back of each person to be tracked. This minimally intrusive setting could be used for example in team sports. The method presented in this chapter shows that the fusion approach significantly improves the tracking performance compared to video-based MPT if video information is missing (*e.g.*, due to occlusions) or ambiguous (*e.g.*, due to similar appearance caused by uniforms).

3.1 Introduction

The tracking-by-detection paradigm is an effective approach for MOT. Instead of directly linking the vast amount of pixels contained in a recording to object identities, the initial detection step drastically reduces the computational effort by grouping the image content to corresponding detections. Yet, relying solely on the output of an object detector is also a major limitation of the paradigm. On the one hand, errors of the detector (see Section 1.3.1) are difficult to compensate for. Frequently missed objects cannot be tracked reliably and misleading false positive detections are difficult to identify and remove. On the other hand, even with perfect detections, it is challenging to create accurate pairwise costs between detections (see Section 1.3.2). When features misinterpret image information misleading signals are produced that potentially lead to wrong tracking results.

To overcome these issues, this chapter presents an HO-MOT formulation that reduces the dependency on the input detections by fusing two signals. To this end, MOT is formulated as a weighted graph labeling problem in Section 3.2. Signals that are assigned the same label correspond to the same object and the corresponding optimization problem is to find the best labeling (see Figure 3.1). A probabilistic model is provided showing that an optimal label solution corresponds to the most likely trajectories.

The formulation fuses signals holistically using identity-dependent pairwise costs which allows taking into account co-occurrences of different objects at the same point of time. The resulting trajectories must be consistent to all input signals, *i.e.*, spatially consistent between different signals and temporally consistent with respect to each signal and across different signals. As all equally labeled measurements contribute to

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

the evaluation of a trajectory, this results in an implicit HO-MOT model.

Finding a global optimal assignment of input signals to labels is difficult as it corresponds to solving an $\mathcal{NP}$-hard binary quadratic problem. A straightforward approach to solve this BQP would be to optimize an equivalent BLP using Theorem 2.50. Yet, in the case of many identities and long sequences, the problem becomes high dimensional so that the corresponding BLP is computationally too expensive and memory demanding to be used.

Instead, a solver is proposed in Section 3.3 that approximates the weighted graph labeling problem. Based on the *Frank-Wolfe* (FW) algorithm, it operates on the continuous relaxation of the BQP. However, using a standard implementation of FW is not sufficient, as the result is often far away from a binary optimal solution. Several crucial modifications are proposed that lead in practice to a much better solution. Consequently, tracking results are substantially improved, as reported in Section 3.4.2. At the same time, the proposed algorithm is much faster than a generic and optimal BQP or BLP solver.

The subsequent sections of this chapter then demonstrate advantages of signal fusion over traditional approaches in two settings:

(i) Section 3.4 proposes to perform MPT by fusing detections that stem from two detectors, namely head and person detections. Even if a person is partially occluded, his or her head is often still visible and can thus be detected (see Figure 1.13) which helps to localize and track persons. At the same time, the fusion facilitates the identification of false positive person detections so that the precision of tracking results can be increased. Consequently, the combination of the complementary inputs signals leads to substantial improvements over the traditional tracking-by-detection approach. Moreover, the resulting tracking method, which we call *Frank-Wolfe Tracker* (FWT), won[2] the MOT 2017 Tracking challenge at the CVPR 2017 against competing approaches.

(ii) Section 3.5 proposes to perform MPT by combining different modalities: a video recording with body-worn *Inertial Measurement Unit*s (IMUs). The fusion reduces the dependence on detection boxes, as motion cues are provided by the IMUs at all times, even if a person is not localized in an image. This allows robust handling of issues such as misleading pairwise costs, or missing and false positive detections. Relating both modalities is challenging, though, as persons walking similarly within a small time window cause similar IMU signals. We term the task of simultaneously assigning detections in a video to IMU devices and object trajectories *Video Inertial Multiple People Tracking* (VIMPT). To tackle VIMPT, we propose IMU device-specific unary costs that link derived orientations of an IMU with regressed orientations of a person using a neural network with a perspective correction method, and pairwise costs that connect accelerations from an IMU with displacements of detection boxes. The fusion formulation together with the device-specific costs, which are essentially independent of the outward appearance of a person, enable to resolve ambiguities in a video signal as well as in the IMU signals.

Once detections in a video are associated with an IMU device, intermediate positions are reconstructed in a post-processing step using a proposed method that jointly optimizes

---

[2] `https://motchallenge.net/MOT17_results_2017_07_26.html`

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

the agreements to both readings. We recorded a challenging VIMPT dataset [121, 122] containing video and IMU recordings together with ground truth annotations. Compared to video-based MOT methods, the proposed VIMPT tracker, which we call *Video Inertial Tracker* (VIT), improves the IDF1 score by a large margin if the video information is ambiguous. In addition, the setup allows to automatically and accurately recover the identity corresponding to a computed trajectory in terms of the IMU device ID.

In summary, this chapter presents methods to improve the tracking accuracy by reducing the dependence on potentially erroneous input detections. This is achieved by:

- A formulation to fuse multiple signals for HO-MOT together with a novel solver for the underlying $\mathcal{NP}$-hard binary linear optimization problem that achieves near-optimal results in practice.
- An MPT method (FWT) that combines people detections with head detections, which enables to identify false positive people detections and track heavily occluded persons.
- Fusing video data with readings from IMUs holistically and globally optimal. The corresponding tracker VIT uses novel device-specific unary and pairwise costs that link orientation and motion information from both modalities. VIT simultaneously tracks and identifies persons in a video. Missing detections are robustly reconstructed using a novel fusion procedure that combines video information from computed trajectories with corresponding IMU signals.

Overall, both fusion approaches lead to substantial improvements over the traditional tracking-by-detection approach. We note that the proposed fusion approach has been successfully applied in further settings, *e.g.*, by augmenting joint detections [101] or for 3D human pose reconstruction [118].

3.2 Signal Fusion as Weighted Graph Labeling Problem

By formulating MOT as a weighted graph labeling problem, two signals can be fused holistically. All connections between equally labeled detections are taken into account so that it corresponds to an implicit HO-MOT method. While it shares similarities to correlation clustering [52, 54, 63, 70], the proposed method uses an efficient labeling formulation that avoids exponential growth of the constraints.

This section presents the weighted graph labeling formulation in terms of a BQP and analyzes its properties. A proof is presented showing that the problem is $\mathcal{NP}$-hard. A probabilistic model for the label weights is provided such that minimal label costs correspond to most likely trajectories. In the case of few identities in a recording, the problem can be solved using an optimal BQP solver [165]. For larger instances, an approximate solver is proposed in Section 3.3.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

3.2.1 Related work

Limiting the input of a tracker to a single detector has clearly several drawbacks since much of the information of the image is not considered, thereby potentially ignoring semi-occluded objects. In recent literature, several works have started incorporating different image features for the task of MPT. Some works use supervoxels as input for tracking, obtaining a silhouette of the pedestrian as a byproduct. Chen *et al.* [189] perform optimization via greedy propagation, while Milan *et al.* [190] formulate supervoxel labeling as a conditional random field.

Several works use dense point tracks [77, 191, 192] or the *Kanade–Lucas–Tomasi feature tracker* (KLT) [193, 194] together with detections to improve tracking performance. Benfold *et al.* [183] propose to track corner features using KLT to obtain a motion model between detections. Fragkiadaki *et al.* [195] tackle MPT by clustering dense feature tracks, and further combine it with detection-based tracklets in a two-step approach [196].

Chari *et al.* [103] track pedestrians using a BQP to fuse head and body detections by modeling non-maxima suppression as well as overlap consistency between features. In contrast to our proposed model, only co-occurrences of active features are considered, while we directly model the grouping of features to different persons; this ensures consistency within each cluster over long time periods. Also, in the extension [197] to motion segmentation using superpixels, the per-person consistency is not considered.

3.2.2 Data association model for signal fusion

The data association between two signals is modeled as a weighted graph labeling problem. To this end, let $(\mathcal{V}, \mathcal{E})$ be the undirected complete graph on the vertex set $\mathcal{V}$ which is composed of all input signals. Given an upper bound on the number of objects to be tracked, denoted as n_{obj}, we consider labelings of the vertices with associated weights (or costs): A map $\mathbf{c} \in \mathbb{R}^{\mathcal{V}\times[n_{\text{obj}}]}$ defines for each node a cost value depending on the assigned label. Between any pair of nodes, $\mathbf{Q} \in \mathbb{R}^{\mathcal{V}\times[n_{\text{obj}}]\times\mathcal{V}\times[n_{\text{obj}}]}$ defines a cost value depending on the assigned labels to both nodes. To consider the functions $\mathbf{c}$ and $\mathbf{Q}$ as vectors, we denote by $[v \rightharpoonup k]$ an index in the corresponding vector representation that represents the assignment of node v to label k. The graph comprising all weighted labels is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{c}, \mathbf{Q}, n_{\text{obj}})$. For each node $v \in \mathcal{V}$, the label weight $c_{[v \rightharpoonup k]} = \mathbf{c}(v, k) \in \mathbb{R}$ reflects the likelihood of v belonging to object k. Node label weights are also called *unary costs*. Likewise, the label weight for an edge $e = \{v, u\} \in \mathcal{E}$ given by $q_{[v \rightharpoonup k],[u \rightharpoonup k']} = \mathbf{Q}(v, k, u, k') \in \mathbb{R}$ measures the consistency for simultaneously assigning node v to label k and node u to label k'. Edge label weights are also called *pairwise costs*. As the graph is undirected, we set $q_{[v \rightharpoonup k],[u \rightharpoonup k']} = q_{[u \rightharpoonup k'],[v \rightharpoonup k]}$ for all $\{u, v\} \in \mathcal{E}$ and for all $k, k' \in [n_{\text{obj}}]$. In the case of $k = k'$, the cost value $q_{[v \rightharpoonup k],[u \rightharpoonup k]}$ reflects how likely v and u belong to the same trajectory. For $k \neq k'$, $q_{[v \rightharpoonup k],[u \rightharpoonup k']}$ reflects consistencies for co-occurrences of different objects.

We introduce indicator variables $x_{[v \rightharpoonup k]}$ which take value 1 if signal $v \in \mathcal{V}$ is assigned to label $k \in [n_{\text{obj}}]$, and 0 otherwise. The indicator variables are stacked in a vector $\mathbf{x} \in \{0, 1\}^{n_{\text{nod}} n_{\text{obj}}}$. Then, the basic idea is to assign a cost value to each hypothesis and

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

select for a sequence the assignments that minimize the total costs, corresponding to the most likely trajectories. For arbitrary $\mathbf{x} \in [0,1]^{n_{\mathrm{nod}} n_{\mathrm{obj}}}$, we associate with $\mathbf{x}$ unary costs given by

$$\mathrm{un}_{\mathcal{G}}(\mathbf{x}) \coloneqq \sum_{\substack{v \in \mathcal{V} \\ k \in [n_{\mathrm{obj}}]}} c_{[v \rightharpoonup k]} x_{[v \rightharpoonup k]} \,, \tag{3.1}$$

and pairwise costs given by

$$\mathrm{pa}_{\mathcal{G}}(\mathbf{x}) \coloneqq \sum_{\substack{\{v,u\} \in \mathcal{E} \\ k,k' \in [n_{\mathrm{obj}}]}} q_{[v \rightharpoonup k],[u \rightharpoonup k']} x_{[v \rightharpoonup k]} x_{[u \rightharpoonup k']} \,. \tag{3.2}$$

These costs are combined to the cost function

$$f_{\mathcal{G}}(\mathbf{x}) \coloneqq \mathrm{un}_{\mathcal{G}}(\mathbf{x}) + \mathrm{pa}_{\mathcal{G}}(\mathbf{x}) \,. \tag{3.3}$$

We define for a video sequence of length n_{R} by $\mathcal{V}_{\mathfrak{f}}$ the set of nodes $v \in \mathcal{V}$ captured at frame $\mathfrak{f} \in [n_{\mathrm{R}}]$ and $\mathfrak{f}_v \coloneqq \mathfrak{f}$ if $v \in \mathcal{V}_{\mathfrak{f}}$. Finally, $n_{\mathrm{nod}} = |\mathcal{V}|$ denotes the number of nodes in $\mathcal{G}$.

Signal fusion via global optimization. For a graph $\mathcal{G}$ as defined above, the data association problem, which fuses multiple input signals, is then given in terms of the following optimization problem.

$$\begin{aligned} \min_{\mathbf{x} \in [0,1]^{n_{\mathrm{nod}} n_{\mathrm{obj}}}} \quad & f_{\mathcal{G}}(\mathbf{x}) && (3.4) \\ \text{s. t.} \quad & x_{[v \rightharpoonup k]} \in \{0,1\}, \quad \forall v \in \mathcal{V}, \forall k \in [n_{\mathrm{obj}}] \,, && (3.5) \\ & \sum_{k \in [n_{\mathrm{obj}}]} x_{[v \rightharpoonup k]} \leq 1, \quad \forall v \in \mathcal{V} \,. && (3.6) \end{aligned}$$

Constraints (3.5) guarantee indicator variables, and constraints (3.6) ensure that each node is assigned to at most one object. We denote the feasibility set of problem (3.4) by $\mathrm{P}_{\mathcal{B}}(\mathcal{G})$ and the polyhedron of the continuous relaxation, *i.e.*, using only constraints (3.6), by $\mathrm{P}(\mathcal{G})$.

Conceptually, $\mathbf{x} \in \mathrm{P}_{\mathcal{B}}(\mathcal{G})$ encodes a clustering such that grouped nodes belong to the same object. In particular, this enables having multiple signals describing an object at the same time, which can be used as a fusion mechanism. The benefits of such an approach are demonstrated in Section 3.4 by fusing multiple detectors. Note also that $\mathrm{P}_{\mathcal{B}}(\mathcal{G})$ has only n_{nod} linear constraints. In contrast to that, commonly used correlation clustering formulations [52, 54, 63, 70] require exponentially many constraints ($\mathcal{O}(2^{n_{\mathrm{nod}}})$) in order to ensure that all edge costs within a trajectory are taken into account.

Integrating non-maxima suppression. A more constrained setting is to require that at each time step, at most one signal (*e.g.*, one detection box) may be assigned to

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

each object, which is ensured by the optimization problem

$$\begin{aligned} \min_{\mathbf{x}\in[0,1]^{n_{\mathrm{nod}}n_{\mathrm{obj}}}} \quad & f_{\mathcal{G}}(\mathbf{x}) && (3.7)\\ \text{s. t.} \quad & x_{[v\rightharpoonup k]} \in \{0,1\}, \quad \forall v \in \mathcal{V}, \forall k \in [n_{\mathrm{obj}}]\,, && (3.8)\\ & \sum_{k\in[n_{\mathrm{obj}}]} x_{[v\rightharpoonup k]} \leq 1, \quad \forall v \in \mathcal{V}\,, && (3.9)\\ & \sum_{v\in\mathcal{V}_{\mathfrak{f}}} x_{[v\rightharpoonup k]} \leq 1, \quad \forall \mathfrak{f} \in [n_{\mathrm{R}}], \forall k \in [n_{\mathrm{obj}}]\,, && (3.10) \end{aligned}$$

where the constraints (3.10) enforce non-maxima suppression. We denote the feasibility set of problem (3.7) by $\mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G})$ and the polyhedron of the continuous relaxation, *i.e.*, without constraints (3.8), by $\mathrm{P}^{\mathrm{NMS}}(\mathcal{G})$.

The feasibility set $\mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G})$ is suitable to fuse two modalities. This is demonstrated in Section 3.5, where the recordings from body-worn IMU devices are fused with video information. Each label corresponds to a unique IMU device, resulting in person-dependent unary and pairwise costs.

Defining the weighted graph labeling problem. We unify both proposed optimization problems, whereby each corresponds to a data association problem:

Definition 3.1. Let $\mathcal{G}$ be a graph defined as above. For $\widetilde{\mathrm{P}}_{\mathcal{B}} \in \{\mathrm{P}_{\mathcal{B}}(\mathcal{G}), \mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G})\}$, the binary quadratic program

$$\mathrm{WGL}(\mathcal{G}, \widetilde{\mathrm{P}}_{\mathcal{B}}) \coloneqq \operatorname*{arg\,min}_{\mathbf{x}\in\widetilde{\mathrm{P}}_{\mathcal{B}}} f_{\mathcal{G}}(\mathbf{x}) \tag{3.11}$$

is a *weighted graph labeling* problem. We shall write $\mathrm{WGL}(\mathcal{G})$ instead of $\mathrm{WGL}(\mathcal{G}, \mathrm{P}_{\mathcal{B}}(\mathcal{G}))$, and $\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$ instead of $\mathrm{WGL}(\mathcal{G}, \mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G}))$. The *relaxed weighted graph labeling* (RWGL) for $\widetilde{\mathrm{P}}_{\mathcal{B}} \in \{\mathrm{P}_{\mathcal{B}}(\mathcal{G}), \mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G})\}$ is given via the corresponding continuous relaxation $\widetilde{\mathrm{P}}_{\mathcal{B}}^{\circ}$ (see Definition 2.44), according to

$$\mathrm{RWGL}(\mathcal{G}, \widetilde{\mathrm{P}}_{\mathcal{B}}) \coloneqq \operatorname*{arg\,min}_{\mathbf{x}\in\widetilde{\mathrm{P}}_{\mathcal{B}}^{\circ}} f_{\mathcal{G}}(\mathbf{x})\,. \tag{3.12}$$

We shall write $\mathrm{RWGL}(\mathcal{G})$ instead of $\mathrm{RWGL}(\mathcal{G}, \mathrm{P}_{\mathcal{B}}(\mathcal{G}))$.

3.2.2.1 Properties of the weighted graph labeling problem

Optimal solutions for linear objectives over $\mathrm{P}_{\mathcal{B}}(\mathcal{G})$. As the next theorem shows, optimizing a linear functional over the LP-relaxation $\mathrm{P}(\mathcal{G})$ results in a binary solution. Consequently, any optimal solution for $\mathrm{P}(\mathcal{G})$ is also an optimal solution for $\mathrm{P}_{\mathcal{B}}(\mathcal{G})$. This fact is exploited in our implementation of the Frank-Wolfe solver, leading to faster computations.

Theorem 3.2. *Given a complete graph $\mathcal{G}$, and let the parameters n_{nod} and n_{obj} be defined as above with respect to $\mathcal{G}$, then the optimal solution of $\min\limits_{\mathbf{x}\in\mathrm{P}(\mathcal{G})} \langle \mathbf{c}, \mathbf{x}\rangle$ is binary for any cost vector $\mathbf{c} \in \mathbb{R}^{n_{\mathrm{nod}}n_{\mathrm{obj}}}$.*

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Proof. To express the constraints (3.6) using a matrix, we assume without loss of generality the variables $\mathbf{x} \in [0,1]^{n_{\text{nod}} n_{\text{obj}}}$ for $\mathcal{V} = \{v_1, \ldots, v_{n_{\text{nod}}}\}$ to be ordered as

$$\mathbf{x} = \left(x_{[v_1 \rightharpoonup 1]}, \ldots, x_{[v_1 \rightharpoonup n_{\text{obj}}]}, \ldots, x_{[v_{n_{\text{nod}}} \rightharpoonup 1]}, \ldots, x_{[v_{n_{\text{nod}}} \rightharpoonup n_{\text{obj}}]} \right)^{\mathsf{T}} . \tag{3.13}$$

For index $s \in [n_{\text{nod}} n_{\text{obj}}]$, let $s_{\text{label}} \in [n_{\text{obj}}]$ denote the label such that $x_s = x_{[v \rightharpoonup s_{\text{label}}]}$ holds[3], using a corresponding node $v \in \mathcal{V}$.

Let $\mathbf{b} \coloneqq \mathbf{1} \in \{0,1\}^{n_{\text{nod}}}$ and $\mathbf{A} = (a_{rs})_{r \in [n_{\text{nod}}], s \in [n_{\text{nod}} n_{\text{obj}}]}$ such that

$$a_{rs} = \begin{cases} 1 & \text{if } s = [v_{\mathrm{r}} \rightharpoonup s_{\text{label}}], \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

By definition, we have $\mathrm{P}(\mathcal{G}) = \{\mathbf{x} \in [0,1]^{n_{\text{nod}} n_{\text{obj}}} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$. Since $\mathbf{A} \in \{0,1\}^{n_{\text{nod}} \times n_{\text{nod}} n_{\text{obj}}}$, all 1×1 sub-matrices of $\mathbf{A}$ have either their determinant equal to one or to zero. Now let $\mathbf{A}' \in \{0,1\}^{n' \times n'}$ be a square sub-matrix of $\mathbf{A}$. Let $\mathbf{a}' = \mathbf{A}'_{[:,1]}$ denote the first column vector of $\mathbf{A}'$. If $\mathbf{a}' = \mathbf{0}$, then $\det(\mathbf{A}') = 0$. Otherwise, since $\mathbf{A}$ has exactly one non-zero entry per column, there exists a row index $r \in [n']$ such that

$$a'_i = \begin{cases} 1 & \text{if } i = r, \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

We conclude by induction that $\mathbf{A}$ is totally unimodular, using Laplace's expansion formula.

Finally, Theorem 2.46 shows that optimal solutions for linear functionals over $\mathrm{P}(\mathcal{G})$ must be binary. □

Complexity. Solving $\mathrm{WGL}(\mathcal{G}, \mathrm{P}_{\mathcal{B}})$ for $\mathrm{P}_{\mathcal{B}} \in \{\mathrm{P}_{\mathcal{B}}(\mathcal{G}), \mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G})\}$ is very challenging as both problems are $\mathcal{NP}$-hard, which is proven in the following.

For $\mathrm{P}_{\mathcal{B}}(\mathcal{G})$, we consider the case $q_{[v \rightharpoonup k],[u \rightharpoonup k']} = 0$ and $c_{[v \rightharpoonup k]} = c_{[v \rightharpoonup k']}$ for all $v, u \in \mathcal{V}$ and for all $k, k' \in [n_{\text{obj}}]$. We can thus write $c_v \coloneqq c_{[v \rightharpoonup k]}$ for all $v \in \mathcal{V}$ and for all $k \in [n_{\text{obj}}]$. We define $y_v \coloneqq \sum_{k \in [n_{\text{obj}}]} x_{[v \rightharpoonup k]}$ for all $v \in \mathcal{V}$. The constraints (3.6) and (3.5) then ensure that $y_v \in \{0,1\}$ for all $v \in \mathcal{V}$. The objective function then simplifies to

$$f_{\mathcal{G}}(\mathbf{x}) = \sum_{v \in \mathcal{V}} c_v \left(\sum_{k \in [n_{\text{obj}}]} x_{[v \rightharpoonup k]} \right) = \sum_{v \in \mathcal{V}} c_v y_v \,, \tag{3.16}$$

which poses an unconstrained binary linear program and is thus $\mathcal{NP}$-hard, according to Theorem 2.43. It follows that $\mathrm{WGL}(\mathcal{G}, \mathrm{P}_{\mathcal{B}}(\mathcal{G}))$ is $\mathcal{NP}$-hard as well.

For $\mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G})$ and the special case $|\mathcal{V}_{\mathfrak{f}}| = 1$ for all $\mathfrak{f} \in [n_{\mathrm{R}}]$, we have $\mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G}) = \mathrm{P}_{\mathcal{B}}(\mathcal{G})$. Again, the problem can be transformed into an unconstrained binary linear program. Accordingly, the more general problem $\mathrm{WGL}(\mathcal{G}, \mathrm{P}_{\mathcal{B}}^{\mathrm{NMS}}(\mathcal{G}))$ is also $\mathcal{NP}$-hard.

Note also that the solution space grows exponentially with the upper bound on the number of objects (n_{obj}) and the number of detections (n_{nod}).

---

[3] The label is computed as $s_{\text{label}} = s - \left\lfloor \frac{s-1}{n_{\text{obj}}} \right\rfloor n_{\text{obj}}$.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Probabilistic model. To ensure that an optimal labeling corresponds to the most likely trajectories, a probabilistic model on the selection of nodes and edges on a graph can be defined in terms of a Bayesian network, which is introduced as in Tang *et al.* [54]: Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph. Consider random variables X and Y, whose realizations $\mathbf{x} \in \{0,1\}^{\mathcal{V}}$ and $\mathbf{y} \in \{0,1\}^{\mathcal{E}}$ correspond to indicator variables. Given a finite index set J and node $v \in \mathcal{V}$, consider a random variable Φ_v, whose realization $\phi_v \in \mathbb{R}^J$ is a vector of features for node v. Accordingly, given a finite index set I and edge $e \in \mathcal{E}$, we consider a random variable Ψ_e with realization $\psi_e \in \mathbb{R}^I$ of features for edge e. Furthermore, we consider the random variables Θ and Θ' with realizations $\theta \in \mathbb{R}^J$ and $\theta' \in \mathbb{R}^I$ of model parameters, respectively. To model the constraint set, a random variable $\mathcal{Z}$ is considered with realization $Z \subset \{0,1\}^{\mathcal{V} \cup \mathcal{E}}$.

Then, according to a Bayesian model, the probability of the indicator variables $\mathbf{x}, \mathbf{y}$, and model parameters θ, θ' given the features ϕ, ψ and constraint set Z factorizes to

$$\begin{aligned} \mathsf{P}(\mathbf{x}, \mathbf{y}, \theta, \theta' \mid \phi, \psi, Z) \propto \mathsf{P}(Z \mid \mathbf{x}, \mathbf{y}) \prod_{v \in \mathcal{V}} \mathsf{P}(x_v \mid \phi_v, \theta) \prod_{j \in J} \mathsf{P}(\theta_j) \cdot \\ \cdot \prod_{e \in \mathcal{E}} \mathsf{P}(y_e \mid \psi_e, \theta') \prod_{i \in I} \mathsf{P}(\theta'_i) \,. \end{aligned} \tag{3.17}$$

The first product constrains $\mathbf{x}$ and $\mathbf{y}$ according to the feasibility set Z:

$$\mathsf{P}(Z \mid \mathbf{x}, \mathbf{y}) \propto \begin{cases} 1, & (\mathbf{x}, \mathbf{y}) \in Z, \\ 0, & \text{otherwise.} \end{cases} \tag{3.18}$$

The second and third term model the classification of each node using a logistic model and a Gaussian prior:

$$\mathsf{P}(x_v = 1 \mid \phi_v, \theta) = \frac{1}{1 + \exp(- \langle \phi_v, \theta \rangle)} \,, \tag{3.19}$$

$$\mathsf{P}(\theta_j) = \mathcal{N}(0, \sigma^2) \,, \tag{3.20}$$

where $\sigma \in \mathbb{R}_{>0}$. The edge classification using the remaining two factors is defined accordingly.

Now using the described probabilistic model, Tang *et al.* [54] derive the cost function such that minimizing that cost function corresponds to maximizing the probability measure:

Theorem 3.3. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph. Moreover, let Z be a feasibility set, let ϕ, ψ be feature vectors, and let θ, θ' be model parameters as defined above with respect to the graph $\mathcal{G}$. Then, a pair $(\mathbf{x}, \mathbf{y})$ with $\mathbf{x} \in \{0,1\}^{\mathcal{V}}$ and $\mathbf{y} \in \{0,1\}^{\mathcal{E}}$ is maximally probable with respect to the measure defined above if and only if it is the solution of the BLP written below with $c_v = - \langle \phi_v, \theta \rangle$ and $q_e = - \langle \psi_e, \theta' \rangle$.*

$$\min_{\substack{\mathbf{x} \in \{0,1\}^{\mathcal{V}} \\ \mathbf{y} \in \{0,1\}^{\mathcal{E}}}} \quad \sum_{v \in \mathcal{V}} c_v x_v + \sum_{e \in \mathcal{E}} q_e y_e \,, \tag{3.21}$$

$$s.\ t. \quad (\mathbf{x}, \mathbf{y}) \in Z \,. \tag{3.22}$$

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Proof. See Tang *et al.* [54]. □

In particular, Theorem 3.3 is applicable to our weighted graph labeling formulation. To see this, we transform graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{c}, \mathbf{Q}, n_{\text{obj}})$ to a graph $\widetilde{\mathcal{G}}$. The vertices of $\widetilde{\mathcal{G}}$ are denoted as $v^{(i)}$ for $v \in \mathcal{V}$ and $i \in [n_{\text{obj}}]$. For $v, u \in \mathcal{V}$, it must hold that $v^{(i)} \neq u^{(j)}$ iff $v \neq u \vee i \neq j$. The edge set $\widetilde{\mathcal{E}}$ of $\widetilde{\mathcal{G}}$ connects any two distinct nodes of $\widetilde{\mathcal{G}}$. Setting $x_{[v \rightharpoonup k]}$ to 1 corresponds to the selection of node $v^{(k)}$. For each product of variables $x_{[v \rightharpoonup k]} x_{[u \rightharpoonup k']}$, a new variable $y_{[v \rightharpoonup k],[u \rightharpoonup k']}$ is introduced such that $y_{[v \rightharpoonup k],[u \rightharpoonup k']} = x_{[v \rightharpoonup k]} x_{[u \rightharpoonup k']}$. If $y_{[v \rightharpoonup k],[u \rightharpoonup k']} = 1$ holds, it corresponds to the selection of a corresponding edge in $\widetilde{\mathcal{G}}$. The equality $y_{[v \rightharpoonup k],[u \rightharpoonup k']} = x_{[v \rightharpoonup k]} x_{[u \rightharpoonup k']}$ can be enforced by linear constraints (see Eq. (2.34)). Costs for the selection of a node and an edge in $\widetilde{\mathcal{G}}$ are given by the respective costs defined by $\mathcal{G}$. It follows that Theorem 3.3 is applicable to the weighted graph labeling problem, using a corresponding constraint set Z.

Thus, we employ logistic regression to learn model parameters $\theta \in \mathbb{R}^J$ and $\theta' \in \mathbb{R}^I$ for the nodes and edges, respectively. In particular, the likelihood that node v corresponds to label k, given the features $\phi_{[v \rightharpoonup k]}$ and model parameters θ is given via

$$p_{[v \rightharpoonup k]} \coloneqq \mathsf{P}(x_{[v \rightharpoonup k]} = 1 \mid \phi_{[v \rightharpoonup k]}, \theta)\,. \tag{3.23}$$

Accordingly, the costs are computed as

$$c_{[v \rightharpoonup k]} = -\langle \phi_{v^{(k)}}, \theta \rangle = \log(\frac{1 - p_{[v \rightharpoonup k]}}{p_{[v \rightharpoonup k]}})\,. \tag{3.24}$$

Likewise,

$$p_{[v \rightharpoonup k],[u \rightharpoonup k']} \coloneqq \mathsf{P}(x_{[v \rightharpoonup k]} = 1 \wedge x_{[u \rightharpoonup k']} = 1 \mid \psi_{[v \rightharpoonup k],[u \rightharpoonup k']}, \theta') \tag{3.25}$$

is the likelihood that node v correspond to label k and node u correspond to label k', given the features $\psi_{[v \rightharpoonup k],[u \rightharpoonup k']}$ and model parameters θ'. The probabilities are transformed to costs, similar to Eq. (3.24).

Nodes that are too far apart in time cannot be compared reliably nor meaningfully. For such edges $\{v, u\}$, we set their corresponding weight to $q_{[v \rightharpoonup k],[u \rightharpoonup k']} \coloneqq 0$ for all $k, k' \in [n_{\text{obj}}]$ which enables incorporating prior knowledge. This strategy effectively sparsifies the graph $\mathcal{G}$ and keeps the proposed approach memory and computationally efficient.

3.3 Frank-Wolfe Optimizer for Weighted Graph Labeling Problems

The proposed weighted graph labeling of Section 3.2 allows fusing two signals holistically using a global HO-MOT formulation. For each signal, all pairwise consistencies within each trajectory are taken into account. Even more, co-occurrences of different objects can be modeled. However, with many objects present in a recording, the corresponding optimization problem is very difficult to solve as it is $\mathcal{NP}$-hard. Consequently, the proposed tracking method relies on finding a good approximation to the solution. To

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

this end, a novel solver based on the Frank-Wolfe algorithm [117] is proposed that approximates the optimal solution well in practice. In particular, a relaxed solution in RWGL($\mathcal{G}$) is computed first, followed by a rounding step[4]. FW is well suited for continuous quadratic problems with linear constraints, as each iteration step involves solving a computationally efficient linear optimization problem. Even more, when the underlying constraints of $\mathrm{P}_{\mathcal{B}}(\mathcal{G})$ are employed, the linear optimization part of Frank-Wolfe can be replaced by an efficient search for the minimal entry in a cost vector. The same principle enables to perform the rounding procedure inside the Frank-Wolfe algorithm efficiently.

When applied to a non-convex problem, as the proposed model $f_{\mathcal{G}}$, the Frank-Wolfe algorithm delivers usually only a local optimum [198]. Hence, simply applying the standard algorithm will result in a solution that is far away from the global optimum.

This work focuses on enhancing the solution of Frank-Wolfe by proposing (i) a regularization of the cost function, (ii) an algebraic and optimal computation of the step size within the solver's algorithm, and (iii) a hierarchical solving scheme that enhances the solution produced by the Frank-Wolfe algorithm. The regularizer prevents the Frank-Wolfe algorithm from falling too quickly into a local optimum. The hierarchical solving scheme gains the improvement by revoking or connecting clusters of the discretized solution. Each iteration results in a smaller problem instance. The procedure is constructed such that subsequent iterations still solve the initial optimization problem. The proposed hierarchical approach is solver agnostic. It can be applied after any approximating algorithm and allows for correcting severe errors introduced by the initial solver.

Experiments in Section 3.4.2 show that the proposed solver provides good solutions close to the estimated bound. As a consequence, the proposed solver achieves state-of-the-art performance on standard MOT datasets, thereby showing the merits of the developed solver and problem formulation. Most importantly, the proposed solver is considerably faster than state-of-the-art BLP and BQP solvers [165] which find in theory a global optimal solution but are not applicable due to memory and computational limitations.

The proposed solver enables to perform optimization directly on the input detections, in contrast to several previous methods [52, 54, 199] which need potentially error-prone initial tracklets.

3.3.1 Related work

Tracking methods that need to solve a BQP have been rare so far due to resulting computational challenges, although many advanced tracking models are naturally expressed as a BQP. For instance, the Markov model [61] can be augmented by an additional detector [103], resulting in a BQP. While this problem can be solved by rewriting the BQP as an equivalent BLP, the experiments in Section 3.4.2 show that this simple trick is not suitable for the more demanding correlation clustering based model we propose, due to the problem size of the BQP. Dehghan *et al.* [200] formulate online tracking via a BQP and solve it using the Frank-Wolfe algorithm, which is also

---

[4] In the context of BLP and BQP, a *rounding procedure* binarizes a solution from a continuous relaxation such that the result is feasible.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

the basis for the developed solver. While the method of Dehghan *et al.* shows good performance, this section presents a hierarchical solving scheme that can be easily integrated into their formulation, thereby potentially further improving their results. Moreover, during the Frank-Wolfe algorithm, the step size for an iterate update has to be computed. An algebraic expression of the optimal step size is presented which is cheap to compute and improves over existing methods [200–202]. Note that the solver may be applied to methods of other fields in computer vision as well, such as person re-identification [201], co-localization [203], or object segmentation [197].

3.3.2 Frank-Wolfe Optimizer for Binary Solutions

We present pseudo-code of the Frank-Wolfe algorithm for the weighted graph labeling problem[5] WGL($\mathcal{G}$) in Algorithm 3.1, which integrates a rounding procedure to obtain feasible binary solutions of $\mathrm{P}_{\mathcal{B}}(\mathcal{G})$.

Algorithm 3.1: Frank-Wolfe algorithm and binarization

```
Input  : Graph $\mathcal{G}$, feasible point $\mathbf{x}^{(0)} \in \mathrm{WGL}(\mathcal{G})$, $\mathsf{IMAX}$, $\epsilon$
Output : Solution vector $\mathbf{x}_{\mathrm{FW}}$, number of iterations $j$

1  $f_{min} \leftarrow \infty$
2  $j \leftarrow -1$
3  repeat
4      $j \leftarrow j + 1$
5      $\mathbf{a}^* \leftarrow \arg\min_{\mathbf{a} \in \mathrm{P}(\mathcal{G})} \nabla f_{\mathcal{G}}(\mathbf{x}^{(j)})\mathbf{a}$
6      $\eta^\star \leftarrow \arg\min_{\eta \in [0,1]} f_{\mathcal{G}}\left(\mathbf{x}^{(j)} + \eta(\mathbf{a}^* - \mathbf{x}^{(j)})\right)$
7      if $f_{\mathcal{G}}(\mathbf{a}^*) < f_{min}$ then
8          $f_{\min} \leftarrow f_{\mathcal{G}}(\mathbf{a}^*)$
9          $\mathbf{x}_{\mathrm{FW}} \leftarrow \mathbf{a}^*$
10     end if
11     $\mathbf{x}_{\mathcal{B}}^{(j)} \leftarrow$ Binarize($\mathbf{x}^{(j)}$)
12     if $f_{\mathcal{G}}(\mathbf{x}_{\mathcal{B}}^{(j)}) < f_{min}$ then
13         $f_{\min} \leftarrow f_{\mathcal{G}}(\mathbf{x}_{\mathcal{B}}^{(j)})$
14         $\mathbf{x}_{\mathrm{FW}} \leftarrow \mathbf{x}_{\mathcal{B}}^{(j)}$
15     end if
16     $\mathbf{x}^{(j+1)} \leftarrow \mathbf{x}^{(j)} + \eta^\star(\mathbf{a}^* - \mathbf{x}^{(j)})$
17 until $[-\nabla f_{\mathcal{G}}(\mathbf{x}^{(j)})(\mathbf{a}^* - \mathbf{x}^{(j)}) < \epsilon] \vee [j > \mathsf{IMAX}]$
```

This is achieved by adding Lines 7-15 to the standard Frank-Wolfe algorithm (compare Algorithm 2.6). The algorithm terminates in case of a small duality gap [204]

$$-\nabla f_{\mathcal{G}}(\mathbf{x}^{(j)})(\mathbf{a}^* - \mathbf{x}^{(j)})\,, \tag{3.26}$$

or if a maximal number of iterations IMAX is exceeded. It produces a binary solution $\mathbf{x}_{\mathrm{FW}}$ that equals either a binarized iterate $\mathbf{x}^{(j)}$ (Lines 11-15) or $\mathbf{a}^*$ (Lines 7-10) as the

---

[5] This section focuses on WGL($\mathcal{G}$). Most results can be adapted for $\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$, *e.g.*, using soft constraints for non-maxima suppression. In Section 3.5 a BLP reformulation of $\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$ is solved globally with acceptable calculation costs as a limited number of persons appear in the recordings.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

constraint matrix corresponding to $\mathrm{P}(\mathcal{G})$ is totally unimodular (see Section 3.2.2.1) so that $\mathbf{a}^*$ is already binary and thus feasible. We refer the interested reader to the literature [203–205] for further details.

Optimizing the first-order Taylor polynomial. As $\mathrm{P}(\mathcal{G})$ is described by a totally unimodular matrix, the optimal solution $\mathbf{a}^*$ in Line 5 of Algorithm 3.1 can be computed efficiently. Since $\mathbf{a}^*$ must be binary, the constraints $\sum_{k\in[n_{\mathrm{obj}}]} a^*_{[v \rightharpoonup k]} \leq 1$ for each $v \in \mathcal{V}$ imply for $\mathbf{w} \coloneqq \nabla f_{\mathcal{G}}\left(\mathbf{x}^{(j)}\right)$ and $k_v \coloneqq \arg\min_{k\in[n_{\mathrm{obj}}]} w_{[v \rightharpoonup k]}$ the optimal solution

$$a^*_{[v \rightharpoonup k]} = \begin{cases} 1, & k_v = k \text{ and } w_{[v \rightharpoonup k]} < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.27}$$

Function `Binarize`. In order to transform a solution $\mathbf{x}^{(j)} \in \mathrm{P}(\mathcal{G})$ into a feasible, binary solution $\mathbf{x}^{(j)}_{\mathcal{B}} \in \mathrm{P}_{\mathcal{B}}(\mathcal{G})$, the function `Binarize` in Line 11 of Algorithm 3.1 discretizes an iterate $\mathbf{x}^{(j)}$ by selecting the closest feasible point $\mathbf{x}^{(j)}_{\mathcal{B}}$ in $\mathrm{P}(\mathcal{G})$ w.r.t. the squared Euclidean distance. It is straightforward to show that

$$\begin{aligned} \mathbf{x}^{(j)}_{\mathcal{B}} &= \operatorname*{arg\,min}_{\mathbf{x}\in \mathrm{P}_{\mathcal{B}}(\mathcal{G})} ||\mathbf{x}^{(j)} - \mathbf{x}||_2^2 && (3.28) \\ &= \operatorname*{arg\,min}_{\mathbf{x}\in \mathrm{P}_{\mathcal{B}}(\mathcal{G})} \left\langle -2\mathbf{x}^{(j)} + \mathbf{1}, \mathbf{x} \right\rangle . && (3.29) \end{aligned}$$

Note that problem (3.29) is linear in $\mathbf{x}$. Thus again, due to the totally unimodular matrix used to describe $\mathrm{P}(\mathcal{G})$, the binarization can be solved efficiently by searching for minimal entries in the cost vector $-2\mathbf{x}^{(j)} + \mathbf{1}$, similar to Eq. (3.27).

3.3.2.1 Memory efficiency

Assuming that the weights are label-independent and no co-occurrences are modeled, we show that the formulation is memory efficient. We thus assume $c_{[v \rightharpoonup k]} = c_{[v \rightharpoonup k']}$, $q_{[v \rightharpoonup k],[u \rightharpoonup k]} = q_{[v \rightharpoonup k'],[u \rightharpoonup k']}$, and $q_{[v \rightharpoonup k],[u \rightharpoonup k']} = 0$ for all $v, u \in \mathcal{V}$ and for all $k, k' \in [n_{\mathrm{obj}}]$ with $k \neq k'$. In this case, we shall write c_v instead of $c_{[v \rightharpoonup k]}$ and $q_{v,u}$ instead of $q_{[v \rightharpoonup k],[u \rightharpoonup k]}$. Moreover, we assume $q_{v,u} = q_{u,v}$ for all $v, u \in \mathcal{V}$. Thus, let $\mathbf{Q}_{\mathrm{pa}} = (q_{v,u})_{v,u\in\mathcal{V}} \in \mathbb{R}^{n_{\mathrm{nod}} \times n_{\mathrm{nod}}}$ and $\mathbf{c}_{\mathrm{un}} = (c_v)_{v\in\mathcal{V}} \in \mathbb{R}^{n_{\mathrm{nod}}}$ be all pairwise and unary costs, respectively. By concatenating n_{obj} times the vector $\mathbf{c}^{\mathsf{T}}_{\mathrm{un}}$, we obtain

$$\widetilde{\mathbf{Q}} = \operatorname{diag}(n_{\mathrm{obj}}, \mathbf{Q}_{\mathrm{pa}}) , \quad \tilde{\mathbf{c}} = (\underbrace{\mathbf{c}^{\mathsf{T}}_{\mathrm{un}}, \ldots, \mathbf{c}^{\mathsf{T}}_{\mathrm{un}}}_{n_{\mathrm{obj}} \text{ times}})^{\mathsf{T}} , \tag{3.30}$$

where $\operatorname{diag}(n_{\mathrm{obj}}, \mathbf{Q}_{\mathrm{pa}})$ is a block diagonal matrix, consisting of exactly n_{obj} identical main diagonal blocks, each given by $\mathbf{Q}_{\mathrm{pa}}$. Note that by construction, $\widetilde{\mathbf{Q}}$ is symmetric. Let $\mathbf{x} \in [0,1]^{n_{\mathrm{nod}} n_{\mathrm{obj}}}$ and $\mathbf{x}_k \coloneqq \mathbf{x}_{[n_{\mathrm{nod}}(k-1)+1\,:\,n_{\mathrm{nod}}k]} \in [0,1]^{n_{\mathrm{nod}}}$ for all $k \in [n_{\mathrm{obj}}]$. The

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

objective function $f_{\mathcal{G}}(\mathbf{x})$ can then be computed as

$$f_{\mathcal{G}}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\intercal}\widetilde{\mathbf{Q}}\mathbf{x} + \langle \tilde{\mathbf{c}}, \mathbf{x} \rangle \tag{3.31}$$

$$= \sum_{k \in [n_{\text{obj}}]} \frac{1}{2}\mathbf{x}_k^{\intercal}\widetilde{\mathbf{Q}}_{\text{pa}}\mathbf{x}_k + \langle \mathbf{c}_{\text{un}}, \mathbf{x}_k \rangle \,. \tag{3.32}$$

By slight abuse of notation, we shall write $\mathbf{Q}$ for matrix $\widetilde{\mathbf{Q}}$ and $\mathbf{c}$ for vector $\tilde{\mathbf{c}}$. Consequently, the potentially huge matrix $\mathbf{Q} \in \mathbb{R}^{n_{\text{nod}} n_{\text{obj}} \times n_{\text{nod}} n_{\text{obj}}}$ and vector $\mathbf{c} \in \mathbb{R}^{n_{\text{nod}} n_{\text{obj}}}$ can be accessed implicitly, using the much smaller matrix $\mathbf{Q}_{\text{pa}}$ and vector $\mathbf{c}_{\text{un}}$. Since $\mathbf{Q}$ is symmetric, the gradient $\nabla f_{\mathcal{G}}\left(\mathbf{x}^{(j)}\right)$ is computed as

$$\nabla f_{\mathcal{G}}\left(\mathbf{x}^{(j)}\right) = (\mathbf{c} + \mathbf{Q}\mathbf{x}^{(j)})^{\intercal} \,. \tag{3.33}$$

Thus, the gradient $\nabla f_{\mathcal{G}}\left(\mathbf{x}^{(j)}\right)$ and the computation of Line 6 of Algorithm 3.1 can be deduced from $\mathbf{Q}_{\text{pa}}$ and $\mathbf{c}_{\text{un}}$, making the application of label-independent costs memory efficient[6].

3.3.2.2 Computing the optimal step size η

The step size $\eta^{\star}$ in Line 6 of Algorithm 3.1 can be computed via line search [206]. However, we derive a new algebraic computation which is faster and global optimal. Let $\mathbf{d}^{(j)} \coloneqq \mathbf{a}^{*} - \mathbf{x}^{(j)}$, $\delta \coloneqq \left\langle \mathbf{d}^{(j)}, \mathbf{Q}\mathbf{d}^{(j)} \right\rangle$, and $\Omega(\eta) \coloneqq f_{\mathcal{G}}(\mathbf{x}^{(j)} + \eta\mathbf{d}^{(j)})$ be given. If $\delta \neq 0$, the only root of the derivative $\frac{d\Omega}{d\eta}$ is given by $\bar{\eta} \coloneqq -\delta^{-1}\nabla f_{\mathcal{G}}(\mathbf{x}^{(j)})\mathbf{d}^{(j)}$. As for the second derivative $\frac{d^2\Omega}{d\eta^2}(\bar{\eta}) = \delta$ holds, the optimal step size $\eta^{\star} = \arg\min_{\eta \in [0,1]} \Omega(\eta)$ is obtained via

$$\eta^{\star} = \begin{cases} \bar{\eta} & \text{if } \delta > 0 \text{ and } \bar{\eta} \in [0,1], \\ 0 & \text{if } (\delta > 0 \text{ and } \bar{\eta} < 0) \text{ or } (\delta < 0 \text{ and } \bar{\eta} > 1), \\ 1 & \text{if } (\delta > 0 \text{ and } \bar{\eta} > 1) \text{ or } (\delta < 0 \text{ and } \bar{\eta} < 0), \\ \underset{\eta \in \{0,1\}}{\arg\min}\, \Omega(\eta) & \text{if } \delta < 0 \text{ and } \bar{\eta} \in (0,1). \end{cases}$$

A line search is needed only in the remaining case $\delta = 0$, making the execution of Line 6 of Algorithm 3.1 very efficient. In contrast to previous works [200, 202], our solution to Line 6 contains all cases that may occur.

3.3.2.3 Regularization of the objective function

Due to the cost function $f_{\mathcal{G}}$ being non-convex, Frank-Wolfe delivers only a local optimum [198]. Consequently, one strategy is to modify $f_{\mathcal{G}}$, enforcing convexity. Given $r \in \mathbb{R}$, we replace the objective function $f_{\mathcal{G}}$ by

$$f_{\mathcal{G},r}(\mathbf{x}) = f_{\mathcal{G}}(\mathbf{x}) + r\sum_{i} (x_i^2 - x_i) \,. \tag{3.34}$$

---

[6] Since $\mathbf{Q}_{\text{pa}}$ is symmetric, storing the upper triangle matrix of $\mathbf{Q}_{\text{pa}}$ is already sufficient.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

For $\mathbf{x} \in \{0,1\}^{n_{\text{nod}} n_{\text{obj}}}$, we have $f_{\mathcal{G},r}(\mathbf{x}) = f_{\mathcal{G}}(\mathbf{x})$. Within $(0,1)$, using $r < 0$ has the effect of pushing the FW algorithm towards discrete solutions, as $-(x_i^2 - x_i)$ has its minimum at 0 and 1. For $r > 0$, we observed better behavior in staying out of local optima, as for a value r sufficiently large, $f_{\mathcal{G},r}$ becomes convex [207–209]. Yet, a too strong convexification leads to degenerated solutions, as also noted by Seguin *et al.* [197]. Replacing each diagonal entry of $\mathbf{Q}$ with the absolute sum of each row makes $\mathbf{Q}$ positive semi-definite and $f_{\mathcal{G},r}$ convex. Therefore, we define $\omega \coloneqq \max\{\sum_b |q_{a,b}| \mid a \in [n_{\text{nod}} n_{\text{obj}}]\}$, set $r_0 = \sqrt{\omega}$ and $r_j = 2^{-j} r_0$. Starting with $r = r_0$, we compute RWGL($\mathcal{G}$) according to $f_{\mathcal{G},r}$ in Algorithm 3.1. Empirically, we observed that the termination of Algorithm 3.1 after a short number of iterations correlates to a bad local optimum caused by a too strong convexification term. Thus, if Algorithm 3.1 terminates in too few iterations (we use 10 as threshold), we reduce the regularization weight by setting $j = j + 1$, $r = r_j$ and re-run Algorithm 3.1 with the updated function $f_{\mathcal{G},r}$. In all our experiments, an appropriate value r was found in at most two function calls of Algorithm 3.1. The entire algorithm is summarized in Algorithm 3.2. In Section 3.4.2, we demonstrate the impact of using the modified cost function with the solver denoted as FW+R.

Algorithm 3.2: Regularized Frank-Wolfe algorithm

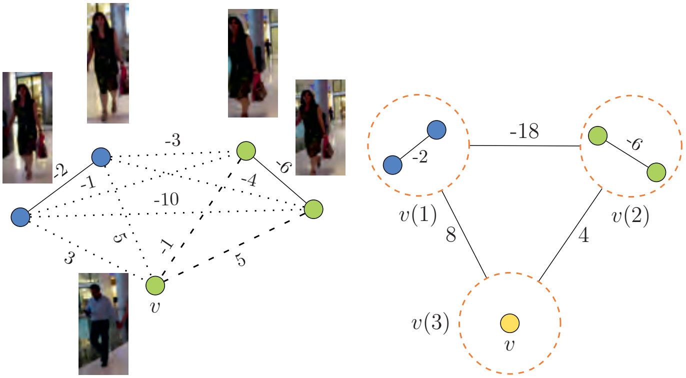
```
Input   : Graph $\mathcal{G}$, feasible point $\mathbf{x}^{(0)} \in \text{WGL}(\mathcal{G})$, IMAX, $\epsilon$
Output  : Solution vector $\mathbf{x}_{\text{FW}}$
$\omega \leftarrow \max\{\sum_{b \in [n_{\text{nod}} n_{\text{obj}}]} |q_{a,b}| \mid a \in [n_{\text{nod}} n_{\text{obj}}]\}$
$r_0 \leftarrow \sqrt{\omega}$
$j \leftarrow -1$
repeat
    $j \leftarrow j + 1$
    $r \leftarrow 2^{-j} r_0$
    $(\mathbf{x}_{\text{FW}}, n_{\text{iter}}) \leftarrow$ Call Algorithm 3.1 using objective function $f_{\mathcal{G},r}$
until $n_{\text{iter}} > 10$
```

3.3.2.4 Hierarchical Solving Scheme

Since FW+R delivers only a local optimum, we propose a new hierarchical solving scheme that enhances the solution of FW+R in each iteration by removing, correcting, and connecting clusters. At the same time, the problem instances become smaller from iteration to iteration so that eventually, the remaining problem instance can be solved to global optimality using an optimal BQP solver. The hierarchical approach is computationally efficient and improves the binary solution w.r.t. the original objective WGL($\mathcal{G}$) in each step.

Compared to other hierarchical approaches [176] that define specific parameter changes in each iteration, the proposed formulation is generic and can be applied to many clustering problems without the need for heuristically set parameter update rules. Pseudo-code of the proposed solving scheme is provided in Algorithm 3.3 and further explained in the following. Note that the algorithm assumes that no co-occurrences are modeled.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a) **(b)**

Figure 3.2: (a) A label result of Algorithm 3.2. Equally colored nodes are grouped, resulting in two clusters. Dotted and dashed edges indicate removed, and selected but wrong connections, respectively. (b) After `LabelCorrection`, node v has been separated, as it has costs 4 to its cluster. Each cluster is represented by a new node. The weighted graph labeling problem is then performed on the contracted graph with updated weights. This allows to further improve the solution. Once the contracted graph is small enough, Gurobi delivers a global optimum. Some parts of the image are taken from the MOT16 dataset [44].

Function `LabelCorrection`: We apply a relabeling strategy that enables correcting errors within the clusters. Those errors may have been introduced due to the rounding or local optimality. Yet, instead of directly resolving wrong assignments, the strategy of the proposed hierarchical solution scheme is to postpone such decisions to a later iteration when more information is available (due to already computed clusters) and when the problem instances are smaller, thus being less prone to approximation errors.

The input for the function at iteration j is the current best labeling $\mathbf{x}^{(j)} \in \mathbb{R}^{n_{\text{nod}} n_{\text{obj}}}$ of $\mathcal{G}$. Initially, we obtain $\mathbf{x}^{(0)}$ using FW+R. Two non-optimal cases are detected:

(i) If a node has positive costs to all of its connected nodes within the same trajectory, separating the node from its trajectory improves the objective value. In more detail, let

$$N(v, \mathbf{x}^{(j)}) := \left\{ u \in N_{\mathcal{G}}(v) \;\middle|\; x^{(j)}_{[v \rightharpoonup k]} = x^{(j)}_{[u \rightharpoonup k]} \; \forall k \in [n_{\text{obj}}] \right\} \tag{3.35}$$

be the set of all adjacent nodes that have the same label as node $v \in \mathcal{V}$. Now any node $v \in \mathcal{V}$ labeled $k \in [n_{\text{obj}}]$ with

$$x_{[v \rightharpoonup k]} \sum_{u \in N(v, \mathbf{x}^{(j)})} q_{[v \rightharpoonup k],[u \rightharpoonup k]} > 0 \tag{3.36}$$

is separated from its trajectory by assigning it to a unique and so-far unused label (see node v in Figure 3.2).

(ii) Bridges within trajectories are verified. To this end, let $e = vu \in \mathcal{E}$ be an edge of equally labeled nodes, *i.e.*, $x^{(j)}_{[v \rightharpoonup k]} = x^{(j)}_{[u \rightharpoonup k]} = 1$ for some $k \in [n_{\text{obj}}]$. Assume that

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

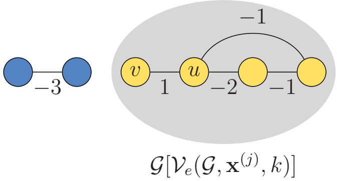

−3
v
u
1
−2
−1
−1
$\mathcal{G}[\mathcal{V}_e(\mathcal{G}, \mathbf{x}^{(j)}, k)]$

Figure 3.3: A label result with a non-optimal assignment at a bridge. For the edge $e = vu$, the induced graph $\mathcal{G}[\mathcal{V}_e(\mathcal{G}, \mathbf{x}^{(j)}, k)]$ is the subgraph within the gray ellipse. Since e is a bridge, assigning node v to a new label improves the objective value.

$\mathfrak{f}_v < \mathfrak{f}_u$ holds. We consider the set

$$\mathcal{V}_{\leq e}(\mathcal{G}, \mathbf{x}^{(j)}, k) \coloneqq \{w \in \mathcal{V} \mid \mathfrak{f}_w \leq \mathfrak{f}_v, x^{(j)}_{[w \rightharpoonup k]} = x^{(j)}_{[v \rightharpoonup k]} = 1\} \tag{3.37}$$

of all nodes labeled k up to frame $\mathfrak{f}_v$ and the set

$$\mathcal{V}_e(\mathcal{G}, \mathbf{x}^{(j)}, k) \coloneqq \{u' \in \mathcal{V} \mid x^{(j)}_{[u' \rightharpoonup k]} = x^{(j)}_{[u \rightharpoonup k]} = 1\} \tag{3.38}$$

of all nodes labeled k. Thus $\mathcal{G}[\mathcal{V}_e(\mathcal{G}, \mathbf{x}^{(j)}, k)]$ contains all nodes and edges corresponding to the trajectory of the object labeled k. Now if e is a bridge in $\mathcal{G}[\mathcal{V}_e(\mathcal{G}, \mathbf{x}^{(j)}, k)]$ and $q_{[v \rightharpoonup k],[u \rightharpoonup k]} > 0$ holds, the objective value improves if all nodes $v \in \mathcal{V}_{\leq e}(\mathcal{G}, \mathbf{x}^{(j)}, k)$ are assigned a so-far unused label k'. An example of this case is shown in Figure 3.3.

Finally, a node $v \in \mathcal{V}$ not selected by the current solution $\mathbf{x}^{(j)}$ is assigned a new and so-far unused label so that node v might be connected to other nodes in a later iteration. Thus if $x^{(j)}_{[v \rightharpoonup k]} = 0$ for all $k \in [n_{\text{obj}}]$, then we set $x^{(j)}_{[v \rightharpoonup k']} = 1$ for a new label k'.

`LabelCorrection` thus results in n'_{obj} used labels and returns corresponding indicator variables as $\mathbf{x}_{\text{corrected}} \in \{0,1\}^{n_{\text{nod}} n'_{\text{obj}}}$.

Function `GraphContraction`: A graph contraction is performed to group all equally labeled nodes and to compute the weighted graph labeling problem on the contracted graph, which is simpler to solve.

To this end, let $v(k)$ be the set comprising all nodes labeled k, according to $\mathbf{x}_{\text{corrected}}$. A *contracted graph* $\underline{\mathcal{G}}$ is build by using these sets as nodes: The vertices are given by $\underline{\mathcal{V}} \coloneqq \{v(k) \mid k \in [n_{\text{obj}}]\}$, and the edge set $\underline{\mathcal{E}}$ is formed by connecting any two different vertices of $\underline{\mathcal{V}}$. The number of nodes is reduced to $\underline{n}_{\text{nod}} \coloneqq |\underline{\mathcal{V}}| \leq |\mathcal{V}|$.

The label costs of $\underline{\mathcal{G}}$ are derived from corresponding costs of $\mathcal{G}$: Given node $v(k) \in \underline{\mathcal{V}}$, which groups all nodes of $\mathcal{G}$ labeled k, we define the unary costs for assigning node $v(k)$ the label k' as

$$\underline{c}_{[v(k) \rightharpoonup k']} \coloneqq \sum_{u \in v(k)} c_{[u \rightharpoonup k']} + \sum_{\substack{\{u,w\} \in \mathcal{E} \\ u,w \in v(k)}} q_{[u \rightharpoonup k'],[w \rightharpoonup k']} \tag{3.39}$$

and pairwise costs for $v_1 \coloneqq v(k_1)$ and $v_2 \coloneqq v(k_2)$ labeled k and k', respectively, as

$$\underline{q}_{[v_1 \rightharpoonup k],[v_2 \rightharpoonup k']} \coloneqq \sum_{\substack{\{w_1,w_2\} \in \mathcal{E} \\ w_1 \in v_1, w_2 \in v_2}} q_{[w_1 \rightharpoonup k],[w_2 \rightharpoonup k']} \,. \tag{3.40}$$

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

The corresponding weighted graph labeling problem on the contracted graph is then defined by $\underline{\mathcal{G}} = (\underline{\mathcal{V}}, \underline{\mathcal{E}}, \underline{\mathbf{c}}, \underline{\mathbf{Q}}, \underline{n}_{\text{nod}})$.

Function `GraphLabeling`: The weighted graph labeling problem is solved on the contracted graph $\underline{\mathcal{G}}$, which is usually much smaller than the initial graph $\mathcal{G}$ so that solving the problem becomes easier. It needs to be assured that (i) the computed solution on the contracted graph is not worse than the current best solution $\mathbf{x}^{(j)}$ on the initial graph $\mathcal{G}$ and (ii) that any solution on the contracted graph can be translated back to a labeling of the initial graph (which is performed by the subsequent function `ExpandGraph`).

By assigning each selected node of $\underline{\mathcal{G}}$ a unique label, *i.e.*, each node $v(k)$ of $\underline{\mathcal{G}}$ with $\exists k' \in [n_{\text{obj}}] : x_{[v(k) \rightharpoonup k']} = 1$, we obtain a label vector $\mathbf{x}_{\text{disjoint}} \in \{0,1\}^{\underline{n}_{\text{nod}} \underline{n}_{\text{nod}}}$ such that $f_{\underline{\mathcal{G}}}(\mathbf{x}_{\text{disjoint}})$ sums up only the unary costs (3.39) defined on $\underline{\mathcal{G}}$. Due to the error correction by the function `LabelCorrection`, it follows that

$$f_{\underline{\mathcal{G}}}(\mathbf{x}_{\text{disjoint}}) \leq f_{\mathcal{G}}(\mathbf{x}^{(j)})\,. \tag{3.41}$$

Thus, the function `GraphLabeling` returns a labeling that is not worse than $\mathbf{x}^{(j)}$. Consequently, solving $\text{WGL}(\underline{\mathcal{G}})$ results in a solution $\mathbf{x}^{\star} \in \text{P}_{\mathcal{B}}(\underline{\mathcal{G}})$ with

$$f_{\underline{\mathcal{G}}}(\mathbf{x}^{\star}) \leq f_{\underline{\mathcal{G}}}(\mathbf{x}_{\text{disjoint}}) \leq f_{\mathcal{G}}(\mathbf{x}^{(j)})\,. \tag{3.42}$$

Note that the graph contraction usually reduces the dimensionality significantly: There are $\underline{n}_{\text{nod}}$ nodes to be labeled using at most $\underline{n}_{\text{nod}}$ labels and usually $\underline{n}_{\text{nod}} \ll n_{\text{nod}}$. If $\underline{n}_{\text{nod}}$ is small enough, we can solve $\text{WGL}(\underline{\mathcal{G}})$ quickly to optimality using Gurobi [165]. Otherwise we use the FW+R solver with $\mathbf{x}_{\text{disjoint}}$ as initial solution.

Function `ExpandGraph`: The label solution $\mathbf{x}^{\star}$ of $\text{WGL}(\underline{\mathcal{G}})$ is translated back to a feasible solution of $\text{P}_{\mathcal{B}}(\mathcal{G})$: Given a node $v \in \underline{\mathcal{V}}$ and let v be assigned the label k, according to $\mathbf{x}^{\star}$. Then all nodes $u \in v$ contained in v are assigned the same label k. The resulting vector of decision variables is the next iterate $\mathbf{x}^{(j+1)}$ of Algorithm 3.3 (see also Figure 3.2). We conclude that

$$f_{\mathcal{G}}(\mathbf{x}^{(j+1)}) = f_{\underline{\mathcal{G}}}(\mathbf{x}^{\star}) \leq f_{\mathcal{G}}(\mathbf{x}^{(j)})\,, \tag{3.43}$$

so that the hierarchical solving scheme can improve the last available iterate. This makes postponing of unclear decisions reasonable.

The loop of the hierarchical solving scheme is stopped, once the objective value does not further improve.

Function `RejectIsolatedNodes`: Finally, isolated nodes with positive unary costs are removed. Hence, if $v \in \mathcal{V}$ is a node labeled k such that $c_{[v \rightharpoonup k]} > 0$ and $x^{(j)}_{[v \rightharpoonup k']} = 0$ for all $k' \neq k$, the objective value improves by not assigning v to any identity. Accordingly, the decision variable is updated by setting $x^{(j)}_{[v \rightharpoonup k]} = 0$.

The effectiveness of the hierarchical solving scheme, which we shall denote by FW+R+H, is demonstrated in Section 3.4.2.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Algorithm 3.3: Hierarchical solving scheme

```
Input  : Graph $\mathcal{G}$, feasible graph labeling $\mathbf{x}^{(0)}$
Output : Solution vector $\mathbf{x}^{(j)}$
$j \leftarrow 0$
repeat
    $f_{\min} \leftarrow f_{\mathcal{G}}(\mathbf{x}^{(j)})$
    $\mathbf{x}_{\text{corrected}} \leftarrow$ LabelCorrection($\mathbf{x}^{(j)}, \mathcal{G}$)
    $\underline{\mathcal{G}} \leftarrow$ GraphContraction($\mathbf{x}_{\text{corrected}}, \mathcal{G}$)
    $\mathbf{x}^{\star} \leftarrow$ GraphLabeling($WGL(\underline{\mathcal{G}})$)
    $\mathbf{x}^{(j+1)} \leftarrow$ ExpandGraph($\mathbf{x}^{\star}$)
    $j \leftarrow j + 1$
until $f_{\mathcal{G}}(\mathbf{x}^{(j)}) = f_{\min}$
$\mathbf{x}^{(j)} \leftarrow$ RejectIsolatedNodes($\mathbf{x}^{(j)}$)
```

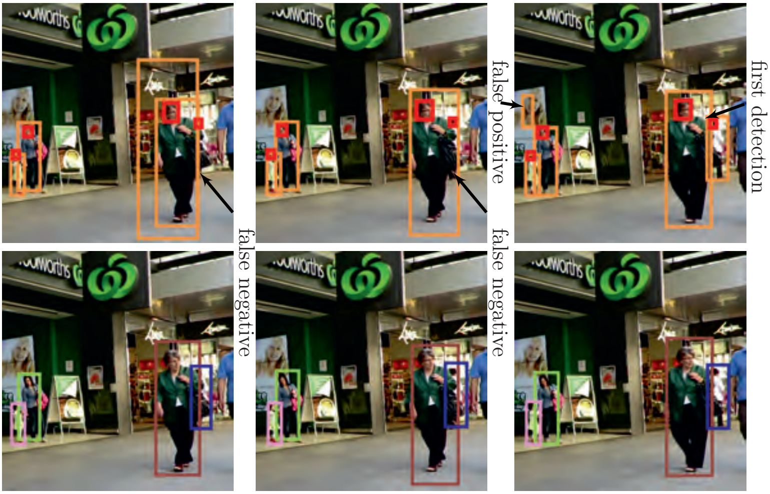


Figure 3.4: MOT16-09 sequence (from left to right: frame 10,12,15). Top row: Body detections (orange) and head detections (red). The people detector misses the person depicted by the arrow until frame 15. Bottom row: The tracking result from the weighted graph labeling formulation by fusing head and people detections. The false positive is removed as it does not have a corresponding head detection. The tracker recovers the heavily occluded pedestrian due to the presence of head detections.

3.4 Multiple People Tracking by Fusing Head and People Detections

Having established the weighted graph labeling formulation in Section 3.2 and a corresponding solver in Section 3.3, this section demonstrates the benefits of the proposed

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

fusion formulation by incorporating two detector types, namely head and people detections. Since heads have few degrees of freedom (*e.g.*, less non-rigid deformations), training a head detector with high accuracy is feasible. Figure 3.4 illustrates the advantage. While the full-body detector misses the highlighted person (due to the occlusion), the person's head is still visible so that the MOT system localizes and tracks that pedestrian correctly. An analysis of the fusion of head detections with people detections reveals that the best tracking accuracy is achieved using both input sources. The fusion helps especially to remove false positive full-body detections that are not consistent with the head detections and to recover heavily occluded persons. The proposed MOT method is used to evaluate the novel solver of Section 3.3. As the experiments show, the proposed improvements to the Frank-Wolfe algorithm are crucial to obtain solutions very close to the optimum. The new solver significantly improves over standard BQP solvers when applied to the weighted graph labeling problem.

Accordingly, the MOT method produces very accurate tracking results. When the work was published, the tracker ranked 2nd on the MOT16 benchmark and 1st on the MOT17 benchmark, outperforming over 90 trackers.

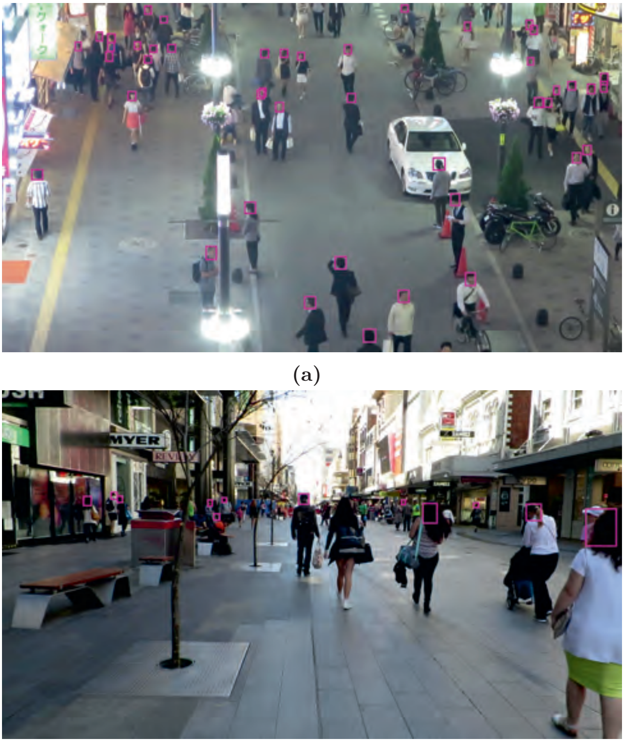
3.4.1 Data association model

Let $\mathcal{V} := \mathcal{V}_\mathrm{P} \cup \mathcal{V}_\mathrm{H}$ be the node set comprising all people detections $\mathcal{V}_\mathrm{P}$ and head detections $\mathcal{V}_\mathrm{H}$ retrieved from a video. The edge set $\mathcal{E} = \mathcal{V}^{(2)}$ connects any two distinct vertices of $\mathcal{V}$. Each node (corresponding to a head or people detection) has assigned label cost values that reflect the probability of the node belonging to a specific person. Similarly, pairwise costs for each edge reflect the probability of the corresponding nodes belonging to a specific person. Co-occurrences between differently labeled nodes are not considered in this setup. Since the persons appearing in a video are not known a priori, label-independent costs $\mathbf{c} \in \mathbb{R}^{\mathcal{V} \times [n_\mathrm{obj}]}$ for nodes and $\mathbf{Q} \in \mathbb{R}^{\mathcal{V} \times [n_\mathrm{obj}] \times \mathcal{V} \times [n_\mathrm{obj}]}$ for the edges are employed, where n_obj denotes a plausible upper bound on the number of persons. The weighted graph labeling formulation WGL($\mathcal{G}$) for $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{c}, \mathbf{Q}, n_\mathrm{obj})$ then allows fusing holistically head detections with people detections for the tracking task, resulting in an implicit HO-MOT model. Logistic regression models (see Section 2.4.2) are trained to form the unary and pairwise costs so that Theorem 3.3 is applicable, ensuring that the most likely trajectories are obtained when solving WGL($\mathcal{G}$). Unary costs are based on the score of each detection. Between nodes, we distinguish spatial and temporal costs, which reflect how likely two detections within the same and between different frames belong to the same person, respectively. Details about the utilized detections and the cost definitions are provided in the following.

Head detections. To obtain accurate head detections, we employ a neural network [210] fine-tuned on the MOT16 training set [44]. Qualitative results are shown in Figure 3.5.

People detections. During training, we use the DPM people detections provided by the MOT16 challenge [44]. For the evaluation, we use the official detections from the MOT16 and MOT17 test set.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a)

(b)

Figure 3.5: Qualitative results of the employed head detector on two sequences of the MOT17 dataset [44].

Relative positioning. In order to obtain meaningful features between differently sized boxes, features have to be formulated respecting the different scales.

To this end, consider a person detection box $\mathbf{d}$ with its corners at the lower left, upper left, and upper right denoted as $\mathfrak{c}_1^{(\mathbf{d})}, \mathfrak{c}_2^{(\mathbf{d})}, \mathfrak{c}_3^{(\mathbf{d})} \in \mathbb{R}^2$, respectively. For a pixel $\mathbf{p} \in \mathbb{R}^2$, we obtain barycentric coordinates $\boldsymbol{\lambda}^{(\mathbf{d},\mathbf{p})} \in \mathbb{R}^3_{\geq 0}$ of $\mathbf{p}$ w.r.t. the corners of $\mathbf{d}$ such that $\mathbf{p} = \sum_{i=1}^{3} \lambda_i^{(\mathbf{d},\mathbf{p})} \mathfrak{c}_i^{(\mathbf{d})}$ and $\sum_{i=1}^{3} \lambda_i^{(\mathbf{d},\mathbf{p})} = 1$. For fixed template box $\mathbf{d}_{\text{tmp}}$, each pixel $\mathbf{p}$ is mapped using its barycentric coordinates to $\mathbf{d}_{\text{tmp}}$:

$$\mathbf{p} \mapsto \sum_{i=1}^{3} \lambda_i^{(\mathbf{d},\mathbf{p})} \mathfrak{c}_i^{\mathbf{d}_{\text{tmp}}}, \tag{3.44}$$

keeping the relative position as in $\mathbf{d}$, see also Figure 3.6. Now, all subsequent distance measurements are computed using the mapped position w.r.t. $\mathbf{d}_{\text{tmp}}$.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

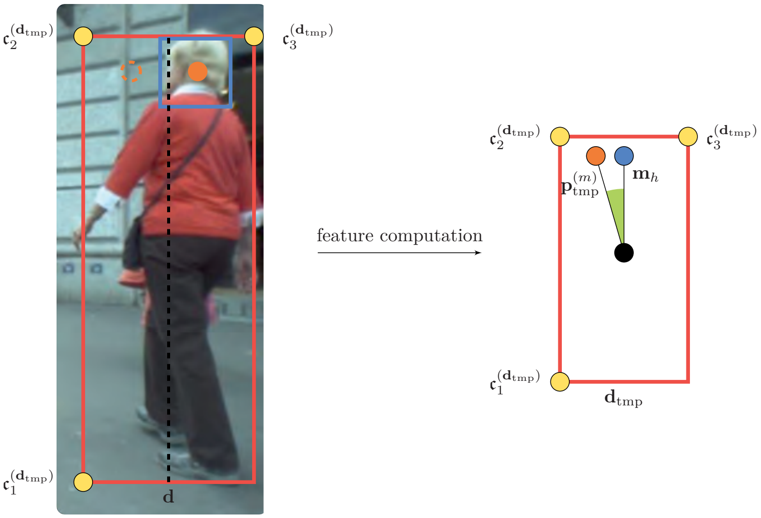


Figure 3.6: Left: Center of head detection (orange) is mirrored to the left half side. Right: Distance and angle computation between expected head position (blue) and observed (mirrored) head position (orange), w.r.t. the template box $\mathbf{d}_{\text{tmp}}$. Some parts of the image are taken from the MOT16 dataset [44].

Spatial costs. We introduce two features that set the position of a head in relation to a person detection.

For a pair of a head and person detection, we mirror the head detection to the left half side of the detection box $\mathbf{d}$, resulting in a position $\mathbf{p}^{(\mathrm{m})}$. This makes the feature robust against different orientations of a person. From the MOT16 training data, the mean relative position $\mathbf{m}_{\mathrm{h}}$ of a head w.r.t. a detection of the same person is learned in terms of the template detection $\mathbf{d}_{\text{tmp}}$. The mean head position is used to compute the feature $\left\|\mathbf{p}_{\text{tmp}}^{(m)} - \mathbf{m}_{\mathrm{h}}\right\|$, measuring the distance between the detected and expected position, where $\mathbf{p}_{\text{tmp}}^{(m)}$ denotes the mirrored head position w.r.t. $\mathbf{d}_{\text{tmp}}$. As second feature, the angle between the expected and detected position of a head is computed with the anchor at the center of the box (see Figure 3.6).

Finally, spatial costs between detections from the same detector type are set to a constant high value.

Temporal costs. Temporal costs are defined via correspondences of pixels between two different frames. DeepMatching (DM) provides such assignments with high accuracy (see Section 2.7.2.1). Given detection boxes v and u, DM samples dm_v and dm_u many pixels in v and u, respectively. Let $\mathrm{co}_{v,u}$ denote the number of correspondences found by DM. (i) To compare two head or people detections, the DM Intersection over union features $\frac{\mathrm{co}_{v,u}}{\mathrm{dm}_v}$, $\frac{\mathrm{co}_{v,u}}{\mathrm{dm}_u}$, and $\frac{\mathrm{co}_{v,u}}{0.5(\mathrm{dm}_v+\mathrm{dm}_u)}$ are used, similar to Tang *et al.* [63]. (ii) As head detections are significantly smaller than people detections, only the temporal head to

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

person feature $\frac{\mathrm{co}_{v,u}}{\mathrm{dm}_v}$ is used, where v denotes a head detection and u a people detection. (iii) From the MOT16 training data, the mean ratios ξ^w_m and ξ^h_m between a head and person detection w.r.t. width and height, respectively, are computed if both belong to the same person. Finally, the features $\left\|\xi^w_\mathrm{m} - \xi^w_{(\mathbf{d},\mathbf{d}')}\right\|$ and $\left\|\xi^h_\mathrm{m} - \xi^h_{(\mathbf{d},\mathbf{d}')}\right\|$ measure the deviation from the expected ratios, using the observed ratios $\xi^w_{(\mathbf{d},\mathbf{d}')}$ and $\xi^h_{(\mathbf{d},\mathbf{d}')}$ w.r.t. width and height, respectively, given a pair of a people and a head detection $(\mathbf{d}, \mathbf{d}')$.

3.4.2 Experimental results

Finally, the proposed solver of Section 3.3 is analyzed in several experiments using the data association model of the last subsection. We call the resulting tracking method *Frank-Wolfe Tracker* (FWT). For the assessment, we report the gain both in speed as well as in tracking performance by the proposed solver. Next, the impact of fusing different detectors on the tracking performance is investigated using the training sequences of the challenging MOT16 benchmark. In the last experiment, the performance of FWT on the test set of the benchmarks MOT16 and MOT17 is presented, showing that the approach achieves state-of-the-art performance compared to competing methods at the time the work was published [55].

3.4.2.1 Implementation Details

In our implementation, we set the temporal costs of two nodes being more than 9 frames apart to zero. The maximal number of labels n_obj is fixed to 70. We process a sequence in batches containing no more than 1800 nodes. We stop the Frank-Wolfe iterations of Algorithm 3.1 in case the duality gap is below 10^{-4} or 750 iterations are reached.

Post-Processing The fusion approach enables the reconstruction of the positions of a person missed by the person detector. Given a pair $(\mathbf{d}_\mathrm{person}, \mathbf{d}_\mathrm{head})$ consisting of assigned person and head detections, the displacement vector from the center of $\mathbf{d}_\mathrm{head}$ to the center of $\mathbf{d}_\mathrm{person}$ and the ratios $\frac{w_\mathrm{head}}{w_\mathrm{person}}$ and $\frac{h_\mathrm{head}}{h_\mathrm{person}}$ of the widths and heads of both detections are computed. If a trajectory is missing a person detection at frame $\mathfrak{f}$ but has assigned head and person detections for frame $\mathfrak{f}'$ with $|\mathfrak{f} - \mathfrak{f}'| \leq \triangle$ and $\mathfrak{f}'$ is closest to $\mathfrak{f}$ with this property, the detection box is reconstructed using the displacement vector, width and height ratios from frame $\mathfrak{f}'$. To avoid error propagation, FWT uses $\triangle = 2$, which was optimized on the MOT16 training set.

A trajectory γ that does not contain any assigned head detection often collects false positive detections, *e.g.*, because the people detector consistently signals the presence of a person at the same background object (*e.g.*, a traffic light). Such a case cannot be recognized using pairwise features as the grouping results in a spatio-temporal consistent trajectory.

Thus, we remove a trajectory γ containing no head detection if the length of γ is at least $\triangle_\mathrm{FP}$ and all people detections have at least the box height h_min. In the experiments, $\triangle_\mathrm{FP}$ is set to 5 and h_min is set to 110 pixels, which was optimized on the training data.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 3.1: Solver comparison: While the proposed solver quickly terminates, Gurobi is not able to finish after 1000 seconds. Entries in brackets denote the results of Gurobi after 1000 seconds.

Method	Iters ↓	Time[sec] ↓	Obj Value ↓	MOTA [%] ↑
FW	**16**	**0.7**	-3060	14.2
FW+R	676	27	-5481	26.8
FW+R+H	-	27+0.5	**-5925**	**27.5**
Gurobi	-	1000	(-5531)	24.9
Gurobi bound	-	1000	(-5973)	-

3.4.2.2 Frank-Wolfe Optimization

The first experiment analyzes the impact of the proposed modifications on the Frank-Wolfe solver. To this end, we choose a representative batch of 41 frames from the most challenging MOT16-13 training sequence and perform tracking using people detections only. It consists of 403 detections so that there are 28210 decision variables. Table 3.1 shows the number of iterations performed by the solver until the duality gap is below the defined threshold, the runtime, the final objective value of f_G as well as the corresponding MOTA score.

The proposed modification FW+R+H (Algorithm 3.3) improves the objective value considerably compared to the standard Frank-Wolfe algorithm FW (Algorithm 3.1). This naturally translates to almost double MOTA accuracy, 14.2% *vs.* 27.5%. Note also that the objective value comes very close to the global optimum. The optimal solver Gurobi [165] is still far away from the global optimum after 1000 seconds, while the proposed solver obtains a much better objective value after only 27.5 seconds. Although Gurobi was not able to compute the global optimum in the given time span, it delivers at each time step a lower bound (Gurobi bound) on the optimal value, showing that the optimal solution to the BQP has an objective value ≥ -5973.

The evolution of the objective value for different solvers is plotted in Figure 3.7. We clearly see when FW stops (blue line), how the regularization FW+R (Algorithm 3.2) improves the objective value by a large margin (green line) and how finally the hierarchical approach FW+R+H (gray line) comes even closer to the estimated lower bound (yellow line), as provided by Gurobi. In contrast, Gurobi (purple line) has a much slower convergence.

To separate the quality of the proposed solver from the detections, we further evaluate the performance on ground truth person detections[7] for 40 frames of each MOT16 training sequence in Table 3.2, where we also report the (relative) duality gap to the optimal solution (GAP). The results show a consistent and huge improvement by the hierarchical concept over FW+R. At the same time, the solutions are close to optimality w.r.t. the objective value and tracking performance. The sequences MOT16-05 and MOT16-11 contain many partial occlusions that make it difficult for the DM features to be correct in any situation, thus resulting in lower tracking scores. However, this

---

[7] To assess only the assignment accuracy, we filter out detections that have less than 15% visibilty.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

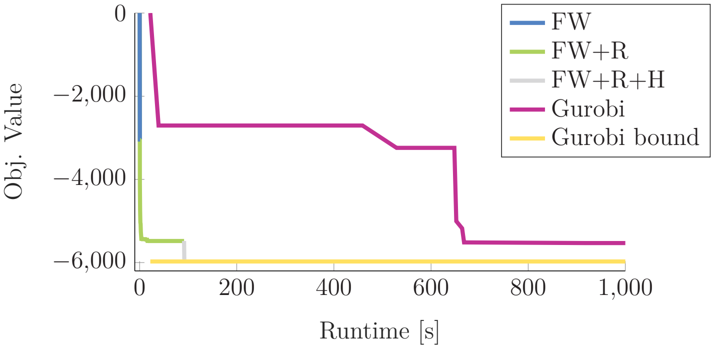


Figure 3.7: Minimization performance of each BQP solver. In addition, a lower bound, as provided by Gurobi, is plotted.

shows that a second type of detections (head detections) is necessary for high-quality tracking results. On the other hand, the solver reaches the perfect result on MOT16-09 (which has far fewer occlusions), thereby justifying the proposed solver.

Table 3.2: FW+R *vs.* FW+R+H, compared on ground truth person detections.

Seq	FW+R					FW+R+H				
	IDF1[%] ↑	IDs ↓	FM ↓	MOTA[%] ↑	GAP[%] ↓	IDF1[%] ↑	IDs ↓	FM ↓	MOTA[%] ↑	GAP[%] ↓
02	87.4	5	1	84.0	6.424	90.9	3	0	90.8	0.428
04	85.0	5	0	73.2	7.506	92.4	0	0	85.8	0.120
05	57.4	10	8	74.2	9.130	70.1	8	7	75.1	0.071
09	80.6	3	0	98.9	5.353	100.0	0	0	100.0	0.000
10	82.0	10	6	80.4	7.410	87.0	7	6	89.4	0.638
11	76.8	13	2	78.2	12.846	89.4	5	3	96.3	0.084
13	87.2	10	2	85.3	10.332	96.3	2	3	96.9	0.434

3.4.2.3 Ablation studies on head and people detections

Next, it is analyzed how the fusion exploits the information provided by the two detectors. Recall that $\mathcal{V}_\mathrm{H}$ and $\mathcal{V}_\mathrm{P}$ denote head and person detections, respectively. In the experiment, people detections are used as provided by the benchmark, while the head detector and the regression models are trained on MOT16 training sequences in a leave-one-out fashion.

Table 3.3 presents the performance of different tracker configurations on the MOT16 training set, depending on different fusion strategies, inputs, and solvers. In order to assess the advantage of a holistic fusion via the weighted graph labeling formulation, we compare against late-fusion, *i.e.*, independently created head and person trajectories that are fused using the proposed solver. The input trajectories are computed by the proposed method (Ours-fusion) and from LP2D [23] (LP2D-fusion) which compares spatial compatibility between detections. We denote the features of LP2D as 2D dist.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

During late-fusion, we use the affinities as defined in Section 3.4.1 but set the spatial and temporal costs between two trajectories originating from the same detector to a constant high value, as the trajectories are already separating the persons (Ours*). Tracker 6 and 7 use the proposed solver and precomputed trajectories from Tracker 3 and 4, showing that the gain using late-fusion is no more than 1.3 percentage points in terms of MOTA. Also, the proposed method (Tracker 4) performs comparably well to Tracker 5 when using people detections and their defined affinities.

We compare the tracking performance to the proposed solver FW+R+H with people detections and head detections (Tracker 10), which is the proposed method FWT. By using the two detectors, FWT significantly improves almost all relevant tracking metrics, justifying the framework and fusion concept. Due to the coupling of head detections with people detections, the number of false positives (FP) is halved. Moreover, persons are tracked more often correctly which results in an increase in the number of mostly tracked (MT) trajectories. Overall, the MOTA score increases by more than 5 percentage points. The quality of the head trajectories (evaluated on self-created head ground truth boxes) is provided by Tracker 2.

We also compare different solver variants. Using another heuristic solver [51] (Tracker 11) performed worse on the fusion than FW+R+H, using exactly the same graph. The comparison Tracker 8 *vs.* Tracker 10 shows a significant improvement on MOT16 train due to the regularizer and the hierarchical step (up to 7.1 percentage points on the MOTA score).

Table 3.3: Ablation experiments on the MOT16 training set.

Detections	Features	Solver	Tracker	MOTA[%] ↑	MT[%] ↑	FP ↓	FN ↓	IDS ↓
$\mathcal{V}_\mathrm{H}$	2D dist	LP2D	1	14.9	**13.5**	14829	50991	472
$\mathcal{V}_\mathrm{H}$	Ours	Ours	2	**16**	**13.5**	**14168**	**50959**	**331**
$\mathcal{V}_\mathrm{P}$	2D dist	LP2D	3	31.7	8.5	**3557**	71332	467
$\mathcal{V}_\mathrm{P}$	Ours	Ours	4	33.0	**14.7**	11949	**61603**	**378**
$\mathcal{V}_\mathrm{P}$	[52]	GMMCP [52]	5	**33.7**	8.9	4053	68675	499
$\mathcal{V}_\mathrm{P} \cup \mathcal{V}_\mathrm{H}$	Ours*	LP2D-fusion	6	33.0	10.4	**3501**	70163	358
$\mathcal{V}_\mathrm{P} \cup \mathcal{V}_\mathrm{H}$	Ours*	Ours-fusion	7	34.2	**16.8**	11852	**60401**	376
$\mathcal{V}_\mathrm{P} \cup \mathcal{V}_\mathrm{H}$	Ours	FW	8	31.1	14.5	5315	69563	1207
$\mathcal{V}_\mathrm{P} \cup \mathcal{V}_\mathrm{H}$	Ours	FW+R	9	33.4	15.9	6497	66238	807
$\mathcal{V}_\mathrm{P} \cup \mathcal{V}_\mathrm{H}$	Ours	Ours	10	**38.2**	16.6	4972	62935	372
$\mathcal{V}_\mathrm{P} \cup \mathcal{V}_\mathrm{H}$	Ours	NLLMPa [51]	11	37.4	16.6	4954	63831	**336**

3.4.2.4 Benchmark Evaluation

Finally, the tracking performance of FWT with people and head detections is evaluated on the MOT16 and MOT17 test set, using the people detections as provided by the respective benchmarks. Table 3.4 shows some of the best performing published trackers as well as the worst performing tracker at the time the method was published [55].

The proposed method FWT creates slightly higher identity switches. We have identified DeepMatching as the main source of the errors. The local image features are prone

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 3.4: Tracking results on the test set of MOT16 and MOT17. Result table retrieved when the work [55] was submitted (04.05.2018).

	Method	Rank↑	MOTA[%]↑	IDF1[%]↑	MT[%]↑	ML[%]↓	FP↓	FN↓	IDS↓
MOT16	LMP [64]	1	**48.8**	51.3	18.2	40.1	6654	86245	481
	Ours	2	47.8	47.8	**19.1**	**38.2**	8886	**85487**	852
	NLLMPa [51]	3	47.6	47.3	17.0	40.4	5844	89093	629
	AMIR [91]	4	47.2	46.3	14.0	41.6	**2681**	92856	774
	NOMT [65]	5	46.4	**53.3**	18.3	41.4	9753	87565	**359**
	GMMCP [52]	15	38.1	35.5	8.6	50.9	6607	105315	937
	DP_NMS [102]	23	26.6	31.2	4.1	67.5	3689	130557	365
MOT17	**Ours**	1	**51.3**	47.6	21.4	**35.2**	24101	247921	2648
	MHT_DAM[211]	2	50.7	47.2	20.8	36.9	**22875**	252889	2314
	EDMT17[212]	3	50.0	**51.3**	**21.6**	36.3	32279	**247297**	**2264**
	GMPHD_KCF[213]	6	30.5	35.7	9.6	41.8	107802	277542	6774

to ID switches if the detections of two persons (people or head detection) share a common image region. When using ground truth detections, much fewer ID changes are generated, compare Table 3.2. This can be explained by the fact that there is much less overlap between ground truth detections belonging to different persons.

However, FWT performs on par with state of the art in terms of tracking accuracy on *MOT16* and sets a new state of the art on *MOT17*. Moreover, the tracker won[8] the MOT 2017 Tracking challenge at the CVPR 2017. The proposed formulation achieves the lowest ML (mostly lost) score within all trackers in both benchmarks, the lowest FN (false negative) score in MOT16, and second lowest FN score in MOT17, showing that the fusion helps to recover more trajectories. Also, the MT score is highest on the MOT16 benchmark and ranks second on the MOT17 benchmark, demonstrating that the method recovers very long trajectories. In contrast, the GMMCP model approach is not able to produce long-term consistent trajectories, possibly due to erroneous initial tracklets that could not be connected[9]. We note that the LMP tracker uses very sophisticated and stable convolutional neural network image features that can reliably link boxes over 200 image frames, thus resulting in a better MOTA and IDF1 score. Finally, visual results on the test set are shown in Figure 3.8.

3.5 Simultaneous Identification and Tracking of Multiple People using Video and IMUs

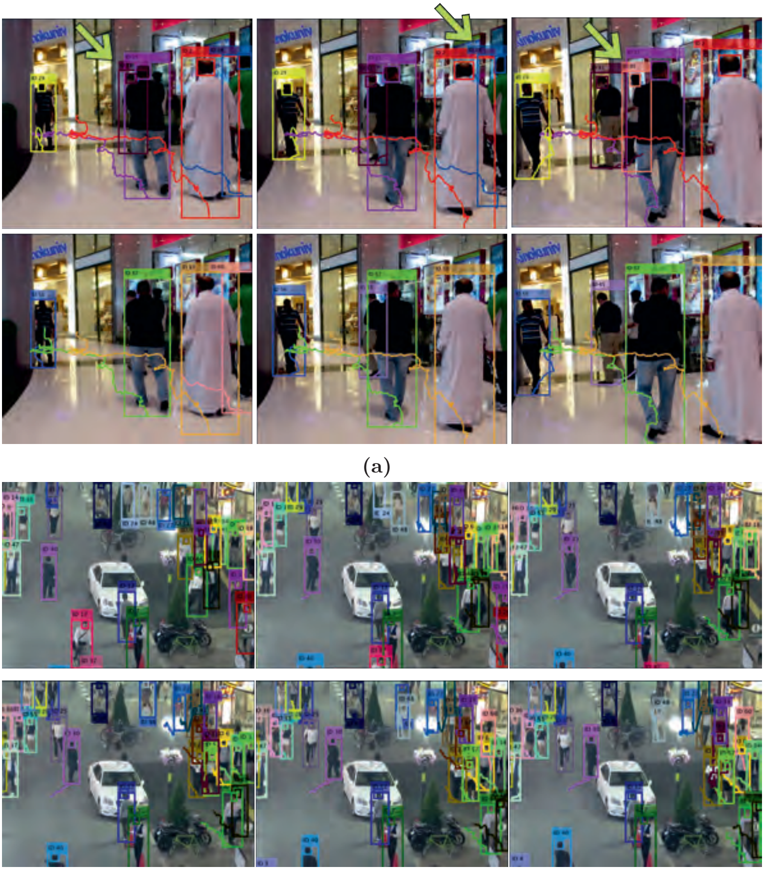
Vision-based MPT methods rely on certain assumptions about the motion and the appearance of the persons to be tracked. A motion model attempts to assign likelihoods to observed person movements. This is very generic and only depends on the corner coordinates of detection boxes. However, as soon as the motion becomes more dynamic, simple motion models [23] are insufficient so that the tracking accuracy degrades. In

---

[8] `https://motchallenge.net/MOT17_results_2017_07_26.html`

[9] We used the official code [52] to produce the results.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a)

(b)

Figure 3.8: (a) Visual comparison on MOT17-12-SDP between the proposed tracker (top row) against the best performing competing method [211] of MOT17 (bottom row). By integration of head detections, heavily occluded persons can be recovered, as highlighted by the arrows. (b) Visual results of FWT on MOT17-03-SDP.

particular, most motion models assume low and constant velocities which holds for pedestrians only within a short temporal window [63]. Also, motion is often measured in 2D, which is ambiguous. Misinterpreting the information thus leads to erroneous trajectories. Another complementary strategy is to model relations between detections based on appearance information. Here, CNN-based feature representations are used to evaluate if two detections show the same person. Recent works have shown very impressive tracking results using only this information [55, 63] or in combination with

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

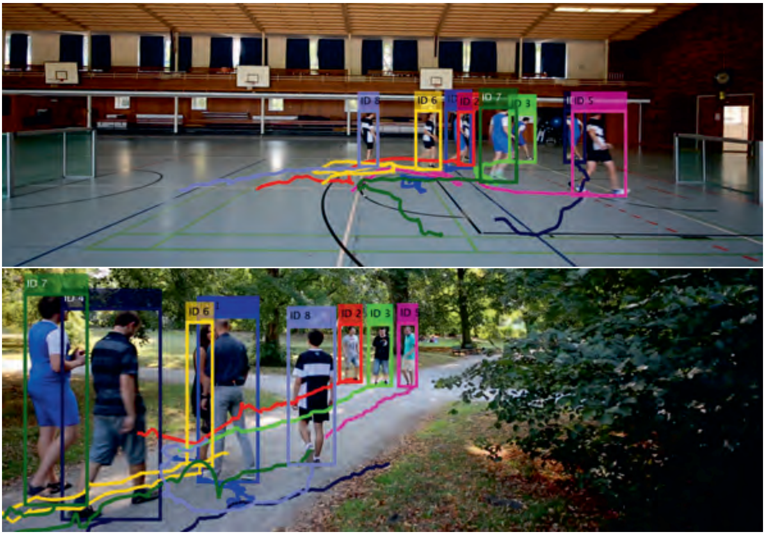


Figure 3.9: Qualitative results obtained by fusing person detections and local motion measurements of body-worn IMUs. Instead of relying on appearance information, the proposed approach enables accurate long-term tracking by finding a globally optimal assignment of detection boxes to IMU devices such that resultant trajectories in the video are consistent with the IMU measurements. This enables to track persons even if they leave and re-enter a scene (see the disconnected trajectories at the lower image border). Since the proposed method automatically assigns each trajectory to an IMU device, ID assignments are consistent across sequences, even if the appearance changes.

motion models [64, 78]. A major advantage of utilizing appearance information over motion models is that it allows relating detections that are temporally far apart. This facilitates the re-identification of people even after long-term occlusions or if they temporarily fall out of the camera view. Despite the enormous progress in obtaining discriminative appearance features, it remains challenging to re-identify persons wearing similar or identical clothing. A prototypical example of such a situation is athlete tracking when team members wear almost identical jerseys. Further challenges arise if low-resolution images are used, when the viewpoint [214] or lighting conditions [215] change, or if people modify their outward appearance. This happens for instance if a person puts on a jacket or opens an umbrella during a recording, but also if people shall be tracked across different recordings, *e.g.*, in a long-term motion study. Then, the assumption of appearance constancy is violated, which degrades tracking accuracy. In summary, computing reliable features between detections is very challenging and error-prone, as also elaborated in Section 1.3.2. Consequently, the task of tracking *and* re-identifying people from visual inputs only is still far from being solved.

To approach this difficult task, we propose to complement visual information from video

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

with motion information using data from body-worn *Inertial Measurement Units* (IMUs). IMUs are small motion sensors measuring[10] local orientation and acceleration.

In particular, we consider a monocular camera view and a single IMU attached to each person to be tracked. Conceptually, the idea is to incorporate local IMU motion measurements in order to disambiguate the assignments of detections to person trajectories. Since IMUs are body-worn, the corresponding motion measurements are unique for each person. Similar to appearance, this property facilitates the re-identification and tracking of persons even after long-term occlusions. Hence, such a tracking approach is predestinated for scenarios where it is possible to equip people with an IMU and appearance is less informative. This could be the case if people wear team jerseys or uniforms or if night-vision is used. Furthermore, the tracking solution provides a unique ID for each trajectory, which corresponds to the associated IMU device. Hence, once the wearer of an IMU device is known, this enables a fully automatic labeling of trajectories to person identities. In contrast, vision-based approaches require manual labeling at this point. Another advantage of combining IMUs and vision for the task of MPT is that inertial sensors enable the reconstruction of people trajectories even if they are occluded or fall out of the camera view. The fusion thus allows to combine the strength of two complementary input sources, and is a promising concept to obtain highly accurate trajectories and tackle tracking and identification holistically, see also Figure 3.9.

The setting allows one (i) to reduce the dependency on artificial motion models (velocities in a video can be related to actual IMU measurements), (ii) to identify persons independent of their outward appearance, and (iii) to automatically assign each trajectory to a person identity.

Incorporating additional sensory input for the task of MPT creates a very different problem setup compared to vision-only tracking methods. In particular, this involves (i) solving the data association problem of detections to trajectories in a video and (ii) simultaneously identifying the corresponding IMU device for each trajectory. Hence, solving this problem requires ensuring consistency within all detections of a trajectory, and at the same time, consistency between each trajectory and the corresponding IMU data. We denote this task as *Video Inertial Multiple People Tracking* (VIMPT).

Even though in VIMPT, motion information is available through IMU measurements, associating these measurements to person detections still poses a very challenging problem. From IMU data alone, it is not possible to generate stable 3D trajectories due to unknown initial states and accumulating drift caused by double integration of acceleration signals [217, 218]. If this was possible, one could easily associate each detection box to the closest IMU trajectory projected to the image. Hence, instead of working on pre-computed IMU trajectories, we have to associate 3D orientation and acceleration measurements to 2D motion information observed in a video. Relating 3D to 2D information under perspective projection is a difficult task on its own. In particular, this requires to relate IMU orientations, which are elements of the 3D rotation group SO(3) [219], to image data being a two-dimensional pixel array. Furthermore, IMU measurements often fit to several people at a time step, and a person wearing an

---

[10] We refer the interested reader to Kok *et al.* [216] for more information on IMUs and corresponding orientation and acceleration signals.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

IMU might be occluded or out of the camera view.

We perform the holistic fusion of video with IMU data by utilizing the proposed weighted graph labeling problem $\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$. The underlying idea is that a global assignment of detection boxes in a video to specific labels representing person identities must be consistent with the measured IMU data. However, this requires a way to measure consistency between video observations and body-worn IMU data. In order to relate IMU orientations to video information, we design a neural network that estimates the orientation of a person within a detection box. Motion cues are incorporated by comparing IMU acceleration measurements to video-based velocities.

To evaluate our proposed tracking approach, we recorded a VIMPT dataset. A special emphasis was put on similar person appearance, heavy occlusions, and non-linear motions. These are situations in which model assumptions implicitly used in vision-based approaches are violated. Since such tracking scenarios are currently missing in standard benchmarks such as DukeMTMC [186] and MOT16 [44], the new dataset could also be valuable for analyzing and improving vision-only approaches.

Evaluations on the recorded dataset reveal that it is possible to achieve reliable tracking results without any motion or appearance model. Moreover, the full system significantly outperforms video-based MOT methods, justifying the VIMPT setup and the tracking method. The resulting VIMPT tracker, which we call *Video Inertial Tracker* (VIT), evaluated on challenging soccer sequences, reduces the number of ID switches (IDS) by more than 70% and doubles the IDF1 score compared to the best performing video-based MPT method. Finally, VIT assigns trajectories to persons on VIMPT without any error.

3.5.1 Related work

Association Weights. The performance of a graph-based tracking approach relies to a great extent on the association weights between detections (or tracklets) that indicate how likely they belong to the same person. As explained in Section 1.3.2, creating discriminative features is challenging, though, as most trackers either employ motion models [22, 23, 89, 95, 97, 220] or appearance models [55, 63, 64, 76–82].

Common to all appearance-based approaches is the assumption of constant and discriminative appearance information. However, these assumptions are violated if persons look identical or change their appearance. Similarly, viewpoint and lighting variations can change the perceived appearance of a person. Motion assumptions are violated for people that are accelerating or moving fast. In contrast, the method presented in this section reduces the dependence on these model assumptions: The orientation of a person is linked with IMU measurements independent of a person's outward appearance and measured accelerations are incorporated into the employed motion model.

Vision and Inertial Sensors. Body-worn inertial sensors provide motion information independent of the visibility of persons. However, it is not possible to recover the 3D person trajectory from IMU information alone [217, 218]. In contrast, using a video signal allows extracting positional information which is complementary to IMU motion

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

information.

Consequently, IMUs have been combined with visual information in many applications, *e.g.*, fusing video and inertial data to stabilize simultaneous localization and mapping (SLAM) [221, 222]. The same modalities have been used to recover human poses [118, 223].

There exist only very few and old works that incorporate IMUs for people tracking in videos. The methods of Jiang *et al.* [224, 225] tackle single person tracking. It requires that a person equipped with an IMU is manually located in the first image in which the person is visible. Then, IMU information is used to recover the trajectory in those situations in which the visual tracker fails. The work of Teixeira *et al.* [226] alternates between multiple people tracking using video information and filtering using IMU information. Some works [227, 228] perform the assignment of trajectories to IMUs once all trajectories have been established.

Instead, the method proposed in this section simultaneously solves the assignment of detections to consistent trajectories and to IMUs, thereby combining the advantages of both sensors. The solutions are obtained from a higher-order data association model using a global optimal solver. We thus argue that the full potential of the two modalities has not been satisfactorily exploited so far.

Other Sensor Modalities. There exist works that incorporate other sensor modalities for MOT, *e.g.*, Camplani *et al.* [229] provide a survey of tracking approaches using RGB-D cameras. However, depth cameras work only indoors and have a limited depth range. Another work [230] integrates video and wireless signals emitted from cell phones. In this setup, the signal quality is used for localization. This is problematic since signal strength heavily depends on unpredictable reflections and absorptions. The proposed VIMPT method does not suffer from these limitations.

Person Identification. Once trajectories are computed, they are used to analyze certain patterns in the motion, *e.g.*, for the purpose of motion segmentation, when point trajectories are used [77, 191], for understanding social behavior of humans [14, 25, 26], or in order to assess the performance of athletes [16–19]. In many of these applications, it is crucial that a particular person is associated with a unique trajectory ID. Ideally, the associated ID is consistent not only throughout a recording but also across different recordings.

Manual labeling of trajectories is a tedious task. Hence, several works focus on automatically obtaining the true identities but consider this as a post-processing step. For instance, determining the labels of computed trajectories can be formulated as a Bayesian network inference problem [231]. By grouping trajectories of the same person together, it allows to improve the ID consistency within a sequence but is not sufficient to re-identify persons across different recordings. By maintaining a database of visual features for each person to be expected in a recording, both tasks (identification within a sequence and across sequences) can be tackled [17]. Trajectories are labeled by employing a conditional random field using the visual features from the database. In contrast to these visual approaches, the proposed VIMPT method allows to simultaneously seek for a solution that is consistent with the video signal and IMU labels so that labeling all

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

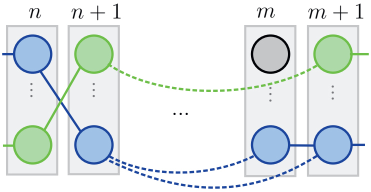


Figure 3.10: Every tracklet is represented by a node in the data association graph of VIT. Each node can be assigned to an IMU device (indicated by color) and is linked to other nodes by short-term edges (solid) and long-term edges (dashed). An edge is activated if the corresponding nodes share the same color. The idea is that every graph color configuration is associated with costs representing the consistency of video information and IMU data. The goal is to find the assignment with minimal costs.

detections is not only a desired task but also helps to obtain accurate tracking results.

3.5.2 Method

The proposed VIMPT method groups detections to short and reliable trajectories (called *tracklets*) in the first step. Then, the goal is to find an optimal assignment of IMU IDs to tracklets such that the resultant trajectories are visually smooth in the video *and* consistent with measured IMU orientations and accelerations, which is illustrated in Figure 3.10.

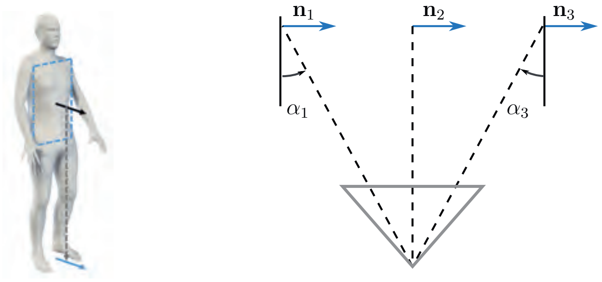
IMU signals are coupled with video information at different conceptual levels: For each potential tracklet to IMU assignment, the person orientation, as seen by the camera, must be consistent with the corresponding IMU orientation. However, orientation consistency alone is very ambiguous. Hence, we also enforce spatio-temporal consistency if two detections are associated with the same ID. Here, we exploit the complementary characteristics of short-term detection box motion features and long-term IMU acceleration features.

3.5.2.1 Data association model

The fusion of video information with the IMU signals is performed in an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{c}, \mathbf{Q}, n_{\text{obj}})$, as introduced in Section 3.2.2. The vertex set $\mathcal{V}$ comprises all tracklets of the entire sequence, and $\mathcal{E}$ is the edge set containing all edges that connect a pair of tracklets. Vertices and edges may obtain a label $k \in [n_{\text{obj}}]$, where n_{obj} denotes the number of persons wearing an IMU.

Solving $\text{WGL}_{\text{NMS}}(\mathcal{G})$ then provides consistent trajectories in the video using tracklets and simultaneously assigns each tracklet to an IMU. By definition, each tracklet is assigned to at most one IMU. Conversely, at any point of time given an IMU, at most one tracklet can be assigned to it.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a) **(b)**

Figure 3.11: (a) We define person orientation in terms of the normal vector of the torso's coronal plane (black arrow) projected to the ground plane (blue arrow). (b) Consider a top view of a person walking on a straight line parallel to the image plane. Even though the person's torso orientation $\mathbf{n}$ is constant (depicted as blue arrows for three distinct positions), the perceived orientation, as seen from the camera, varies. In particular, the perceived orientation differs from the global orientation by an angle α which describes the angle between the depth axis of the camera (straight line) and a vector pointing from the camera center to the person position (dashed line). We use this angle to correct person orientation estimates from the camera in order to relate them to global orientation measurements from body-worn IMUs.

Using short but reliable tracklets has the advantage that more meaningful and more accurate unary and pairwise features can be used. Furthermore, using tracklets significantly reduces the problem size. This enables to solve the weighted graph labeling problem $\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$ to optimality by applying Gurobi [165] on a BLP transformation of $\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$.

Next, the unary and pairwise label costs are described in detail. Specifically, consistency features are introduced that are mapped to costs, as described later in Section 3.5.3.3.

3.5.2.2 Unary Features

In order to measure the likelihood of an assignment hypothesis $x_{[v \rightharpoonup k]}$, the orientation of a person in each detection box of a tracklet v is estimated and compared to the temporally aligned orientation measurements by an IMU k.

In this work, the person orientation $\mathbf{n} \in \mathbb{R}^2$ is defined as the normal vector of the torso's coronal plane projected to the ground plane, as illustrated in Figure 3.11(a). We use the projected normal (instead of the 3D normal) as this comprises fewer degrees of freedom and people usually move in a rather upright pose.

Hence, given the image data $I_\mathbf{d}$ of a detection $\mathbf{d}$, the heading $\hat{\mathbf{n}}_\mathbf{d}$ of the corresponding person needs to be estimated. However, the observed heading in $I_\mathbf{d}$ is affected by the

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

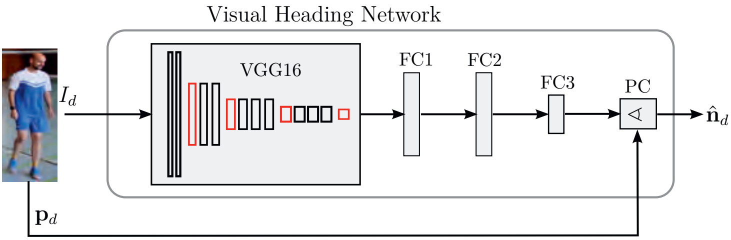


Figure 3.12: The Visual Heading Network predicts the heading $\hat{\mathbf{n}}_{\mathbf{d}}$ of a person using the image data $I_{\mathbf{d}}$ of detection $\mathbf{d}$. Based on the box position $\mathbf{p}_{\mathbf{d}}$, the network performs a *Perspective Correction* (PC) in the last layer. Hyperbolic tangent activation functions are used after each fully-connected layer subsequent to the VGG-16 backbone.

position of the person in the image. To see this, consider a person walking on a straight line parallel to the image plane of a non-moving camera. In a global context, this person has a constant orientation. However, due to perspective effects, the perceived orientation of that person with respect to the viewpoint of the camera is different at every point in the image, see also Figure 3.11(b). We compensate for this using a correction angle derived from the detection box within the image. Let $\alpha_{\mathbf{d}}$ be the angle between the vector defined by the camera center and lower central anchor point $\mathbf{p}_{\mathbf{d}}$, and the depth-axis of the camera. In order to compensate for the perspective influences, the perceived orientation is rotated by $-\alpha_{\mathbf{d}}$, resulting in the prediction $\hat{\mathbf{n}}_{\mathbf{d}}$, compare Figure 3.11(b).

In order to obtain the person heading from image data, we propose to utilize a neural network that provides the mapping $I_{\mathbf{d}} \mapsto \hat{\mathbf{n}}_{\mathbf{d}}$. To this end, VGG-16 [172] pre-trained on ImageNet [130] is extended to regress the heading, which also incorporates the aforementioned *Perspective Correction* (PC) in the last layer. The resulting network is referred to as the *Visual Heading Network* (VHN) in the following. A graphical illustration of the network architecture is depicted in Figure 3.12. In the VIMPT setting, IMUs are consistently placed on the back of each person such that the local sensor's z-axis corresponds to the normal vector of the torso's coronal plane. Hence, the measured torso orientation vector $\mathbf{n}_{k,\mathfrak{f}}$ of IMU k at frame $\mathfrak{f}$ is obtained according to

$$\mathbf{n}_{k,\mathfrak{f}} = \pi(\mathbf{R}_{k,\mathfrak{f}}\mathbf{z}) \,, \tag{3.45}$$

where $\mathbf{z} = (\,0\;0\;1\,)^\intercal$ is the local z-axis vector, $\mathbf{R}_{k,\mathfrak{f}} \in \mathrm{SO}(3)$ is the measured orientation of IMU k at frame $\mathfrak{f}$ mapping the local sensor coordinate frame to the global coordinate frame, and $\pi : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ projects the normal vector to the ground plane. Finally, we measure the deviation of the predicted orientation from the IMU heading vector in terms of the cosine similarity.

In detail, we define the unary orientation feature representing the likelihood of assigning node v to label k as

$$\phi^{(\mathrm{ori})}_{[v \rightharpoonup k]} = \frac{1}{N_v} \sum_{\mathbf{d}\in v} L_{\mathrm{CS}}(\hat{\mathbf{n}}_{\mathbf{d}}, \mathbf{n}_{k,\mathfrak{f}_{\mathbf{d}}}) \,, \tag{3.46}$$

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

where L_{CS} denotes the cosine similarity defined in Eq. (2.19) and N_v denotes the number of detections contained in tracklet v. The orientation feature $\phi^{(\mathrm{ori})}_{[v \rightharpoonup k]}$ measures the average orientation consistency of the tracklet v to the IMU with ID k.

3.5.2.3 Pairwise Features

We define pairwise features that represent the compatibility of two assignment hypotheses $x_{[v \rightharpoonup k]}$ and $x_{[u \rightharpoonup k]}$ with $k \in [n_{\mathrm{obj}}]$. Two nodes v, u are *compatible* if jointly labeling v and u with the same label is reasonable with respect to spatio-temporal aspects.

Spatio-Temporal Features. Within a short temporal window, a person cannot move arbitrarily fast. Hence, compatible tracklet pairs should be spatially close and corresponding detection boxes should be similar in size. We derive corresponding features in the following.

The lower central anchor point $\mathbf{p}_\mathbf{d}$ of a detection $\mathbf{d}$ can be projected to the ground plane, resulting in a position $\mathbf{d}_{\mathrm{plane}} \in \mathbb{R}^2$ in world coordinates. A person detected by $\mathbf{d}$ is thus located near $\mathbf{d}_{plane}$. For detections $\mathbf{d}$ of v and $\mathbf{d}'$ of v', let $\mathbf{v}_{3\mathrm{D}}(\mathbf{d}, \mathbf{d}')$ denote the velocity in 3D from $\mathbf{d}$ to $\mathbf{d}'$ (using the projection) and let $N(v,u)$ be the set of all pairs of detections between v and u. We define a mean velocity feature between nodes via

$$\psi^{(\mathrm{vel})}_{[v \rightharpoonup k],[u \rightharpoonup k]} = \frac{1}{|N(v,u)|} \sum_{(\mathbf{d},\mathbf{d}') \in N(v,u)} \|\mathbf{v}_{3\mathrm{D}}(\mathbf{d}, \mathbf{d}')\|_2 \,. \tag{3.47}$$

Additionally, the heights of detections contained in both tracklets are compared: Let $h_\mathbf{d}$ denote the height of detection box $\mathbf{d}$ in pixels. We define a compatibility measure $\widetilde{\psi}^{(\mathrm{height})}(\mathbf{d}, \mathbf{d}')$ based on the heights of detections $\mathbf{d}$ and $\mathbf{d}'$ according to

$$\widetilde{\psi}^{(\mathrm{height})}(\mathbf{d}, \mathbf{d}') = \widetilde{\psi}^{(\mathrm{time})}(\mathbf{d}, \mathbf{d}') \frac{|h_\mathbf{d} - h_{\mathbf{d}'}|}{\min\{h_\mathbf{d}, h_{\mathbf{d}'}\}} \,, \tag{3.48}$$

where the factor in front of the fraction compensates for the temporal distance between $\mathbf{d}$ and $\mathbf{d}'$, reducing the weight of this comparison with higher temporal distances. In detail, if $\mathbf{d}$ and $\mathbf{d}'$ are s frames apart, we set

$$\widetilde{\psi}^{(\mathrm{time})}(\mathbf{d}, \mathbf{d}') = \frac{1}{\log(c+s)} \,, \tag{3.49}$$

where c is chosen such that $\log(c+1) = 1$. Then, $\widetilde{\psi}^{(\mathrm{time})}(\mathbf{d}, \mathbf{d}') = 1$ holds if the frame distance between $\mathbf{d}$ and $\mathbf{d}'$ is 1, and it increases slowly with bigger frame distances s. Finally, we define our box height feature as

$$\psi^{(\mathrm{height})}_{[v \rightharpoonup k],[u \rightharpoonup k]} = \frac{1}{|N(v,u)|} \sum_{(\mathbf{d},\mathbf{d}') \in N(v,u)} \widetilde{\psi}^{(\mathrm{height})}(\mathbf{d}, \mathbf{d}') \,. \tag{3.50}$$

Both $\psi^{(\mathrm{vel})}$ and $\psi^{(\mathrm{height})}$ are features that are meaningful within short temporal windows. However, this section focuses on sequences in which people get occluded or fall out of the camera view quite often and for longer time periods. Hence, in the following, we utilize acceleration measurements to link hypotheses that cover larger temporal horizons.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Acceleration Feature. Ideally, the ground position $\mathbf{p}_{k,t_1} \in \mathbb{R}^2$ of an IMU k at time t_1 can be recovered by double integration of the corresponding acceleration signal $\mathbf{a}_k$ according to

$$\mathbf{p}_{k,t_1} = \mathbf{p}_{k,t_0} + \mathbf{v}_{k,t_0}(t_1 - t_0) + \int_{t_0}^{t_1} \int_{t_0}^{u} \mathbf{a}_k(s) ds\, du\,, \tag{3.51}$$

where t_0, $\mathbf{p}_{k,t_0}$, and $\mathbf{v}_{k.t_0}$ denote initial time, initial position, and initial velocity, respectively. Please note that $\mathbf{a}_k$ in this case represents the gravity-free acceleration in global coordinates.

Now let $\mathbf{p}_{t_0}$ and $\mathbf{p}_{t_1}$ denote the 2D ground positions of detections $\mathbf{d}$ and $\mathbf{d}'$, respectively. After double integration of the acceleration signal, we can solve Eq. (3.51) for the initial velocity, which we denote as $\mathbf{v}_{\mathrm{IMU}}$. Thus,

$$\mathbf{v}_{\mathrm{IMU}}(\mathbf{d}, \mathbf{d}', k) = \frac{\mathbf{p}_{t_1} - \mathbf{p}_{t_0} - \int_{t_0}^{t_1} \int_{t_0}^{u} \mathbf{a}_k(s) ds\, du}{t_1 - t_0}\,. \tag{3.52}$$

For $\mathbf{d} \in v$, the velocity $\mathbf{v}_{\mathbf{d}}$ of a person at initial time t_0 can be approximated in terms of finite differences, using detections $\mathbf{d}' \in v$ of the same tracklet that are temporally close to $\mathbf{d}$. Hence, for a compatible hypotheses pair of tracklets v and u, the velocity differences

$$\widetilde{\psi}_k^{\text{v-diff}}(\mathbf{d}, \mathbf{d}') = \|\mathbf{v}_{\mathrm{IMU}}(\mathbf{d}, \mathbf{d}', k) - \mathbf{v}_{\mathbf{d}}\|_2 \tag{3.53}$$

should be small for all possible detection pairs $\mathbf{d} \in v$ and $\mathbf{d}' \in u$. We define the acceleration feature as the set of all such differences according to

$$\psi^{(\mathrm{acc})}_{[v \rightharpoonup k],[u \rightharpoonup k]} = \left\{ \widetilde{\psi}_k^{\text{v-diff}}(\mathbf{d}, \mathbf{d}') \mid (\mathbf{d}, \mathbf{d}') \in N(v, u) \right\}\,. \tag{3.54}$$

3.5.2.4 Trajectory reconstruction based on signal fusion

The solution to problem $\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$ assigns each detection box to at most one IMU device. Given these associations, it becomes feasible to reconstruct the trajectory of a person using IMU accelerations accurately. In particular, the positions of a person can be recovered in all frames, even if the person is temporarily occluded or falls out of the camera view. This is a unique advantage of the VIMPT setting.

In the following, we consider a trajectory γ tracking a person with label k and define $\underline{\gamma}$ to be the lower central anchor points of the trajectory projected to the ground plane, described in world coordinates. Moreover, let $\mathfrak{f}_{\mathrm{F}}$ and $\mathfrak{f}_{\mathrm{L}}$ denote the timestamps of the first and last detection assigned to trajectory γ, respectively. Then, the reconstruction method has to recover the locations for all image frames $[\mathfrak{f}_{\mathrm{F}} : \mathfrak{f}_{\mathrm{L}}]$.

Given the correspondences between visual information and an IMU device k, the trajectory γ is improved by seeking for a trajectory that is spatially close to γ while being consistent with the motion information given by IMU sensor k (in terms of acceleration).

Let $\Gamma(\mathfrak{f}_{\mathrm{F}}, \mathfrak{f}_{\mathrm{L}})$ denote the set of all feasible trajectories on the ground plane within the time window $[\mathfrak{f}_{\mathrm{F}} : \mathfrak{f}_{\mathrm{L}}]$. Given the acceleration measurements $\mathbf{a}_k$ from the IMU device k, we seek for a trajectory $\hat{\gamma} \in \Gamma(\mathfrak{f}_{\mathrm{F}}, \mathfrak{f}_{\mathrm{L}})$ that minimizes the following optimization problem:

$$\min_{\hat{\gamma} \in \Gamma(\mathfrak{f}_{\mathrm{F}}, \mathfrak{f}_{\mathrm{L}})} w\, e_{\mathrm{pos}}(\hat{\gamma}, \gamma) + (1 - w)\, e_{\mathrm{acc}}(\hat{\gamma}, \mathbf{a}_k)\,, \tag{3.55}$$

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

where the residual

$$e_{\text{pos}}(\hat{\gamma}, \gamma) = \sum_{\mathfrak{f} \in \text{supp}(\gamma)} \frac{\left\| \hat{\gamma}(\mathfrak{f}) - \underline{\gamma}(\mathfrak{f}) \right\|_2^2}{|\text{supp}(\gamma)|} \tag{3.56}$$

measures the mean squared distance between the estimated trajectory $\hat{\gamma}$ and the detections of trajectory γ. The support of a trajectory γ, denoted as $\text{supp}(\gamma)$, is the set of frames for which γ contains a detection. The residual

$$e_{\text{acc}}(\hat{\gamma}, \mathbf{a}) = \sum_{t \in T_{\text{IMU}}(\mathfrak{f}_F, \mathfrak{f}_T)} \frac{\| \mathbf{a}(t) - \hat{\mathbf{a}}_{\hat{\gamma}}(t) \|_2^2}{|T_{\text{IMU}}(\mathfrak{f}_F, \mathfrak{f}_T)|} \tag{3.57}$$

measures the mean squared distance between the acceleration signal $\mathbf{a}$ given by the IMU signal and the approximated acceleration in the video described by the estimated trajectory $\hat{\gamma}$. Here, the acceleration $\hat{\mathbf{a}}_{\hat{\gamma}}(t)$ of a trajectory $\hat{\gamma}$ at time t is approximated via finite differences:

$$\hat{\mathbf{a}}_{\hat{\gamma}}(t) := \frac{\hat{\gamma}(t - \triangle t_{\text{IMU}}) - 2\hat{\gamma}(t) + \hat{\gamma}(t + \triangle t_{\text{IMU}})}{(\triangle t_{\text{IMU}})^2}\,, \tag{3.58}$$

where $\triangle t_{\text{IMU}}$ is the time distance between consecutive IMU signals. The set $T_{\text{IMU}}(\mathfrak{f}_F, \mathfrak{f}_T)$ denotes the set of timestamps[11] within the first and last detection of γ at which IMU signals exist.

The parameter w can be used to balance the importance of each input channel. The optimization problem (3.55) has a non-linear least squares form. We apply the Levenberg-Marquardt algorithm [232, 233] (see also Section 2.6.5) to obtain the estimated trajectory. The balancing weight is optimized on the training data.

3.5.3 Evaluation

Finally, the tracking approach for the VIMPT setup is evaluated, which we call *Video Inertial Tracker* (VIT). Technical details are provided and the performance is assessed. We evaluate tracking accuracy with respect to several relevant tracking and re-identification metrics and examine the influence of IMU features. In order to demonstrate the advantages of incorporating IMU data, we also compare to vision-based MPT baselines.

3.5.3.1 VIMPT2019 dataset

We created a dataset, called VIMPT2019, to assess the proposed VIMPT method. Our recordings contain challenging sequences captured with a calibrated camera and body-worn IMUs. In the following, an introduction of the dataset and details about the recording procedure are provided, followed by a discussion about challenges of the dataset.

Sequences. The dataset comprises 7 challenging soccer and outdoor recordings. In total, it contains nearly 6500 frames captured with a static camera and 8 IMU-equipped actors in varying clothing styles.

---

[11] Given the framerates of the video and IMU signal, we can evaluate $\hat{\gamma}$ at time t and frame $\mathfrak{f}$.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

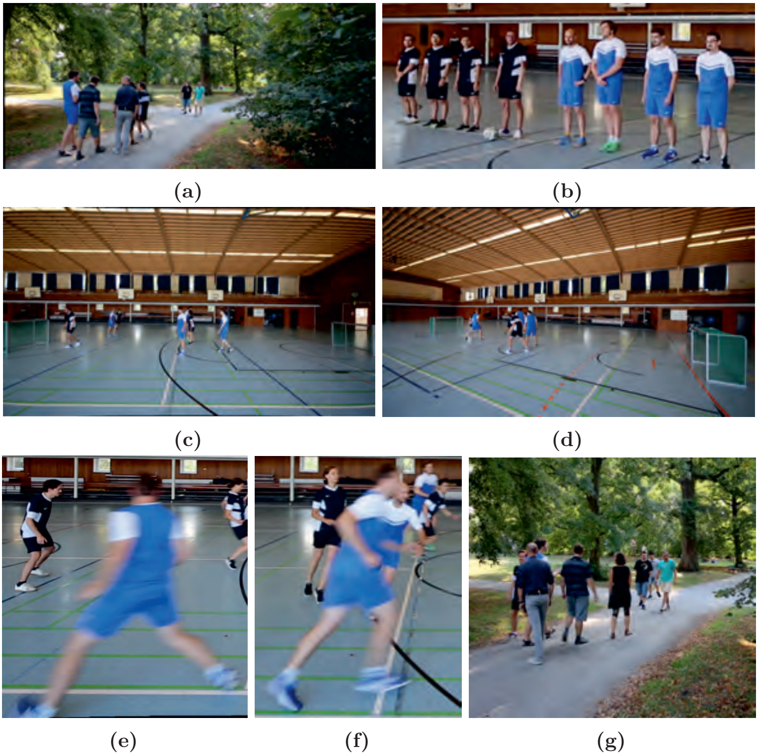


Figure 3.13: First row: Different scenes in the VIMPT dataset. (b) shows very similar person appearances. Second row: Different camera views in the soccer sequences. Third row: Challenges in the VIMPT dataset. (e) Rapid motions and motion blur. (f) Heavy occlusions. (g) Outdoor scene with frequent occlusions.

During soccer recordings, two four-person teams in team jerseys (see Figure 3.13(b)) play soccer in a competitive manner. Consequently, these recordings contain a lot of motion, motion blur, abrupt changes in direction, and occlusions, see Figure 3.13(e)-(f). Hence, tracking challenges arise from non-linear motion and ambiguous appearance information. Moreover, the soccer sequences are captured from two different viewpoints and differ in recorded game situations.

In addition to the soccer recordings, the VIMPT2019 dataset contains an outdoor sequence recorded at a pedestrian crosswalk in a public park (see Figure 3.13(a,g)). Actors walk around in natural apparel and meet regularly for short conversations. This sequence serves as a reference to standard benchmarks such as MOT16/17 [44] and DukeMTMC [186], since it is comparable in terms of motions and scenery. Throughout all sequences, actors regularly leave the field of view and are heavily occluded by other actors.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

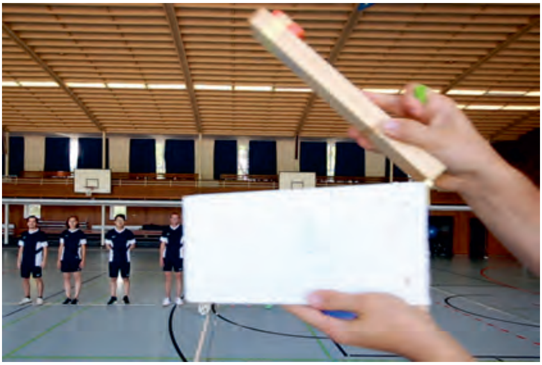


Figure 3.14: Time synchronization using a clapper board.

Camera setup. For all sequences, a calibrated camera has been mounted to a tripod at a height of approximately 1.8m. The videos were captured in landscape mode at 30Hz with 1920 × 1080 spatial resolution, and the camera's extrinsic matrix was calibrated to a fixed reference point in the scene.

Time Synchronization. All wireless IMU devices are automatically synchronized using the recording system of the IMU manufacturer Xsens [234]. For the time synchronization between the IMU devices and a video recording, an additional IMU has been attached on a clapperboard. The clapperboard allows detecting the shut of the clap within the video and the IMU signal (see Figure 3.14).

Detections. The detector FRCNN [73] is trained on MS COCO [235] to generate person detections within all frames of the dataset. For all detections, we compute the corresponding 3D positions using the homography between ground and image plane. In addition, we manually created ground truth detection boxes and labeled them with the corresponding person IDs. Similar to MOT16 [44], we interpolated ground truth detections for occluded persons.

IMU setup. Throughout all sequences, eight persons were equipped with an IMU. Each sensor was attached to a person at hip height. IMU orientation and acceleration were captured at a frame rate of 60Hz. We calibrated the inertial reference coordinate frame to the same reference point as used for the extrinsic camera parameters.

Training, validation and test split. The VIMPT dataset is split into disjoint subsets. One soccer sequence is selected for training and validation of tracker parameters while the residual six sequences are used for testing and evaluation.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 3.5: Characterization of the VIMPT2019 dataset.

Name	Length	Boxes	μ(vis)	σ^2(vis)	Activity	Density
Rec01	1189	7496	0.84	0.10	Football	7.48
Rec02	1148	7428	0.83	0.11	Football	7.90
Rec03	1067	6677	0.77	0.15	Football	7.34
Rec04	653	4384	0.77	0.19	Football	7.65
Rec05	726	4848	0.81	0.11	Football	7.42
Rec06	542	3653	0.79	0.15	Football	7.72
Rec07	1036	6358	0.83	0.11	Walking	6.74

3.5.3.2 Characteristics of the VIMPT2019 dataset

Many different state-of-the-art MPT datasets exist (see Section 2.7.3), each focusing on certain challenges of multiple people tracking. Some concentrate on a wide variety of camera views [44], containing low- to semi-crowded scenes while others focus on long-term tracking [186] with sometimes very crowded scenes and filmed by multiple static cameras. The recordings of VIMPT2019 focus on ambiguous appearance information and non-linear motions.

Quantitative analysis. Table 3.5 lists several characteristics for each sequence of the dataset. The lengths of the sequences range from 540 to nearly 1200 frames, similar to the MOT16/17 sequences. The third column of Table 3.5 shows the number of detections per sequence. *Density* denotes the average number of ground truth detections per frame, indicating the number of people present in the scenes. We further provide a visibility distribution of the dataset. To this end, we compute for each ground truth box $\mathbf{d}$ a visibility score $\mathrm{vis}(\mathbf{d})$. We define a box $\mathbf{d}'$ to partially occlude a box $\mathbf{d}$ with respect to the image coordinates from a camera C, denoted as $\mathbf{d} \lhd_C \mathbf{d}'$, if the boxes $\mathbf{d}$ and $\mathbf{d}'$ have a non-empty intersection and if the lower central anchor point of $\mathbf{d}'$ is lower than the lower central anchor point of $\mathbf{d}$. Then, we compute the relative number of pixels within $\mathbf{d}$ that are not occluded by other ground truth boxes:

$$\mathrm{vis}(\mathbf{d}) = 1 - \frac{\left|\bigcup_{\mathbf{d} \lhd_C \mathbf{d}'} \mathbb{I}(\mathbf{d}') \cap \mathbb{I}(\mathbf{d})\right|}{|\mathbb{I}(\mathbf{d})|}, \tag{3.59}$$

where $\mathbb{I}(\mathbf{d})$ is the set of image pixels contained in box $\mathbf{d}$. The mean visibility score $\mu(vis)$ is provided in Table 3.5 and ranges between 0.79 and 0.84 with a standard deviation $\sigma^2(vis)$ between 0.10 and 0.19.

Characterization of the visual information. We measure the difficulty of distinguishing individuals using visual information. In particular, we compute the Rank-1 and the mAP scores for the MOT16 dataset and the VIMPT2019 dataset using a state-of-the-art re-identification method [236] fine-tuned on DukeMTMC [186]. Additionally, we restrict the gallery and query set to contain the same number of identities, making the evaluations between the different datasets comparable. In particular, we randomly

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 3.6: Comparison between MOT16, VIMPT2019, and its subset (VIMPT2019*) consisting of all football sequences. We compare the difficulty of re-identifying persons using the Rank-1 and mAP metrics. For a motion characterization, we compare the speed distribution of ground-truth trajectories in different datasets: 2D velocities are converted into normalized speed values and aggregated by computing the mean $\overline{v}$ and the variance $\sigma^2(v)$.

Dataset	Rank-1[%] ↑	mAP[%] ↑	$\overline{v}$ [1/s]	$\sigma^2(v)$ [1/s^2]
MOT16	90.3	88.0	0.19	0.06
VIMPT2019	67.4	81.1	0.60	0.34
VIMPT2019*	63.4	78.3	0.63	0.36

sample query and gallery images from the ground truth detections of a randomly selected sequence, obtaining two disjoints sets, and evaluate the metrics. The results shown in Table 3.6, which have been averaged over 100 repetitions, indicate that it is much harder to track people correctly in the VIMPT2019 dataset, as the appearance information is ambiguous due to the worn soccer jerseys.

Characterization of the motions. We characterize the speed distribution of the VIMPT2019 dataset. To this end, we use the ground truth trajectories and compute the average image speed between any two detections $\mathbf{d}_1$ and $\mathbf{d}_2$ belonging to the same trajectory, normalized by the mean box height between $\mathbf{d}_1$ and $\mathbf{d}_2$, in order to compensate for perspective effects. To put the values into perspective, we compare to the MOT16 dataset[12] and present the results in Table 3.6. The evaluations show that VIMPT2019 contains much higher speeds and most importantly, the speeds have a variance, which is more than 5 times higher compared with the MOT16 dataset. Accordingly, it is much more difficult to define a discriminative motion affinity that covers the whole range of plausible movements for the VIMPT2019 dataset.

3.5.3.3 Tracker Parameters

Tracklet generation. The proposed VIMPT method VIT is based on reliable tracklets, which are generated by grouping detections using the method of Pirsiavash *et al.* [102]. In order to avoid error propagation, temporally subsequent detections can only be connected if their intersection over union is above 0.7. The maximal tracklet length is set to 0.5 seconds.

Visual Heading Network. The overall network architecture is depicted in Figure 3.12. It contains the VGG-16 architecture, which is truncated after its last pooling layer. The layers FC1, FC2, and FC3 are fully connected layers with 16, 16, and 2 neurons, respectively. To output an orientation vector n that is within the unit sphere S^1, we use hyperbolic tangent activation functions. Note that VGG-16 has been trained on ImageNet with an invariance for horizontal flipping [172]. To undo this, we train

---

[12] We use all training sequences filmed by a static camera.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

the layers FC1, FC2, and FC3 together with the last convolutional layer of VGG-16 while keeping the weights of all other layers fixed. During training, we add dropout layers [151] with $p = 0.3$ between the fully connected layers to avoid overfitting. Using dropout makes the prediction more robust against clutter within a detection box (*e.g.*, other objects or body parts of other people) so that the neural network is forced to predict the orientation given any arbitrary body part. The network was trained using RMSprop [237] for 250 epochs with a learning rate of 10^{-4} and a batch size of 16. Input images of detection boxes were scaled to 250×675.

Finally, the network weights $\mathbf{W}$ of VHN are learned by maximizing the average cosine similarity between predicted and ground truth heading vector

$$\frac{1}{|D|} \sum_{\mathbf{d} \in D} L_{\mathrm{CS}}(\hat{\mathbf{n}}_{\mathbf{d}}(\mathbf{W}), \mathbf{n}_{\mathbf{d}}) , \tag{3.60}$$

given all ground truth detections $\mathbf{d} \in D$ and corresponding IMU heading vectors $\mathbf{n}_{\mathbf{d}}$ of the VIMPT training sequence.

Graph edge settings. A weighted edge $e \in \mathcal{E}$ between two nodes $v, u \in \mathcal{V}$ is created in the following cases. If the smallest frame distance between all detections of v and u is within $\{1, \ldots, 12\}$, a short-term edge with costs derived from box features is established. If the smallest frame distance is between $\{13, \ldots, 150\}$, long-term edges associated with costs derived from acceleration features are established.

Feature to cost mapping. In order to transform unary and pairwise features to costs, two concepts are used. For orientation and box features, we apply a logistic regression model, which is trained using the ground truth trajectories of the VIMPT training sequence.

We observed that this does not work satisfactorily for the acceleration feature as noise in the 3D position estimates destroys much of the expressiveness of this feature. Instead, we use a threshold δ to indicate if two hypotheses are highly incompatible. Hence, we assign a high constant cost to an edge between nodes v and u labeled k if $\min \psi^{(\mathrm{acc})}_{[v \rightharpoonup k],[u \rightharpoonup k]} > \delta$.

3.5.3.4 Tracking Evaluation

The goal of the proposed VIMPT method is to track IMU-equipped persons in a video accurately. A perfect tracking result is achieved if the assignment of person-specific IDs to corresponding tracklets is consistent throughout the whole tracking sequence.

We evaluate tracking performance by assessing assignment consistency in terms of ID metrics IDP, IDR, and IDF1 but also report the event-based metric MOTA, see Section 2.7.4. Note that the computations of false positives and false negatives within MOTA are based solely on detection existence for each frame independently. Moreover, ID switches have a low weight in MOTA since the number of false positives and false negatives are usually several orders of magnitude higher. In contrast to that, IDF1 evaluates false positives (resp. negatives) *and* also verifies that the person IDs are correct. In particular, the IDF1 score incorporates the longest coverage of each ground

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 3.7: Tracking accuracy on the soccer sequences.

Tracker	IDP[%] ↑	IDR[%] ↑	IDS ↓	MOTA[%] ↑	IDF1[%] ↑
DeepCC [78]	26.3	27.9	395	11.8	27.1
DeepSORT [97]	49.6	42.4	193	77.1	45.8
FWT [55]	29.7	26.7	489	71.6	28.1
VIT	**93.6**	**90.1**	**44**	**86.1**	**91.8**

Table 3.8: Tracking accuracy on the outdoor recording.

Tracker	IDP[%] ↑	IDR[%] ↑	IDS ↓	MOTA[%] ↑	IDF1[%] ↑
DeepCC [78]	55.4	57.0	47	67.3	56.2
DeepSORT [97]	53.4	48.0	28	**83.5**	50.5
FWT [55]	39.1	36.3	66	82.4	37.6
VIT	**89.5**	**78.5**	**22**	81.8	**88.5**

truth trajectory by exactly one computed trajectory. Thus we consider IDF1 as the more meaningful metric for the VIMPT task. Yet, MOTA is a well-known metric for MPT and it enables to put the tracking results into the context of other works.

Tracking accuracy. We report tracking accuracy of our approach, denoted as Video Inertial Tracker (VIT), on the VIMPT dataset in the bottom row of Table 3.7 and 3.8. For the challenging soccer sequences, VIT achieves a very high IDF1 score of 91.8%. Hence, for all IMU-equipped persons, we find and correctly assign almost all corresponding tracklets in a video. This works even though the motions are very dynamic and people get occluded or temporarily leave the field of view. The overall good tracking performance is also supported by the other metrics. Additionally, we obtain almost identical scores for the park sequence, which contains less dynamic motions but is comparable in terms of people visibility. This proves that our approach is not limited to sports tracking but generalizes to other scenarios too. Note that the FRCNN input detections perform very well on the VIMPT sequences so that there are very few false positives or false negatives, which is why the margin on the MOTA score is very small between the different trackers.

Comparison to vision-based methods. We apply three vision-based MPT methods[13] to the VIMPT dataset, namely the MPT method FWT of Section 3.4, DeepSORT [97], and DeepCC [78]. DeepSORT is an online tracker using a sophisticated motion model, and DeepCC focuses on re-identifying persons across different cameras. These approaches have in common that they rely on appearance to establish affinities between detection boxes. The parameters of these trackers were used as provided by the respective authors.

Within the soccer sequences, all team players wear identical jerseys; hence, the appear-

---

[13] The methods represented the state of the art when the VIMPT method was published [121].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

ance information is very ambiguous. The tracking results shown in Table 3.7 validate that this is very challenging for all considered trackers. Respective IDF1 scores vary between 27.1% and 45.8%. In contrast, by using IMU information, VIT can double the IDF1 score to 91.8%. The other metrics show the same trend, and also the MOTA score of VIT is approximately 9 percentage points higher compared to appearance-based approaches. However, a comparison of VIT to the vision-only[14] trackers is not completely fair. They use different sensor modalities and also, the number of tracked people is not fixed for the vision-only approaches. However, the results demonstrate the advantages of incorporating IMU data if appearance is ambiguous and also validate that the proposed fusion algorithm works accurately.

Interestingly, for the park sequence, where people have discriminative appearance, our proposed tracker is on par with the other trackers when MOTA is considered. In contrast, the IDF1 score of VIT is still higher, indicating that people specific trajectories are recovered more accurately by VIT. Finally, comparing Table 3.7 with Table 3.8 indicates that the biggest gain over vision-based tracking systems is achieved when the appearance information is not discriminative. In those cases, the IMU devices compensate for the misleading information.

Reconstruction effectiveness. We analyze the benefit of having IMU information available when recovering missing detections. To this end, we compute the tracking performance using the reconstruction method introduced in Section 3.5.2.4 and analyze its robustness on the VIMPT2019 dataset. Moreover, we compare the results against vision-only based linear interpolation.

To this end, we remove randomly selected input detections and compute the reconstruction accuracy in terms of tracking metrics. We repeat the computations 10 times and plot the mean value of each metric in Figure 3.15. The experiment shows a significantly better performance when the reconstruction method using IMU information is used, with a difference of more than 10 percentage points in terms of IDF1 and more than 20 percentage points of MOTA when all input detections are used. With more detections removed, the improvement over video-based interpolation increases significantly.

Accordingly, using a fusion of video and IMU, the tracking approach becomes more robust against occlusions and less dependent on appearance information. When 10% of the input detections are removed, the performance is almost the same as using the entire detection set. Even when 30% of the detections are artificially suppressed, the performance drop is still acceptable and clearly better than using the vision-only based linear interpolation, *e.g.*, the IDF1 score drops from 91% to 75% for the method of Section 3.5.2.4 while it drops from 78% to about 28% using linear interpolation. Even more dramatically, the MOTA score drops from 60% to -26% using linear interpolation, while the fusion method yields a drop from 86% to 65%. This experiment clearly demonstrates the benefits of VIT over vision-based approaches. Objects may be occluded or even out of view and can still be tracked reliably.

---

[14] Previous works [226, 227] on multiple people tracking combining video with IMUs essentially perform trajectory filtering using IMU data, employ inaccurate detectors, and simple data association models with only local considerations. A comparison to them is therefore neither far nor meaningful.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Influence of IMU features. In order to investigate the influence of orientation and acceleration measurements on the tracking result, we report tracking accuracy of five tracker variants: Ori, Acc+Ori, VT, VT+Acc, and VT+Ori. We evaluate all trackers on the VIMPT dataset and show the results in Table 3.9.

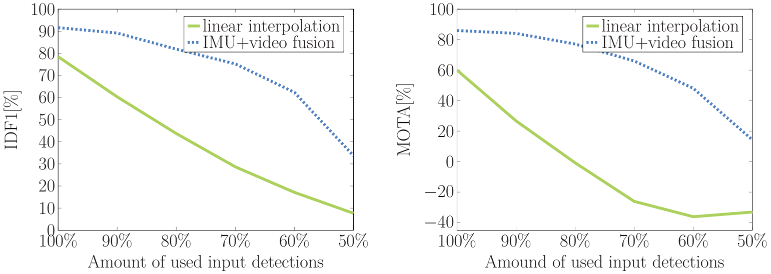
Ori incorporates only the heading information from the VHN network, setting all other costs to zero. It reaches an IDF1 score of 48.7% and a MOTA score of 45.1%, which is already 53% of the overall MOTA and IDF1 performance of VIT. Given that no temporal information has been used for this tracker, the results show the effectiveness of using orientation predictions. However, note that distinguishing people by means of their heading vectors poses a very challenging problem. Especially for the VIMPT dataset, heading predictions have to be very accurate in order to differentiate people successfully. This is due to the soccer sequences, where soccer players are often oriented similarly to follow the game ball. Consequently, these sequences provoke a high number of IDS. In this sense, the soccer recordings can be seen as a worst-case test setup for IMU-based tracking.

The tracker VT uses only box features with all costs related to IMU data set to zero. It obtains an IDF1 score of 38.1%, which is approximately 58% worse compared to VIT. VT+Acc extends VT by taking the acceleration feature into account and applying the reconstruction method of Section 3.5.2.4 based on the IMU's acceleration signal. This helps to recover more detections and to form consistent trajectories. Accordingly, the MOTA score increases by 25% and the IDF1 score by about 20% compared to VT. However, recall that due to measurement noise, the impact of the acceleration feature on the data association had to be weakened to a simple thresholding rule. In contrast, incorporating orientation information to VT, denoted as VT+Ori, leads to a significant increase in tracking accuracy, yielding an IDF1 score of 76.4%. Hence, the orientation consistency in combination with the simple motion model is key to disambiguate tracklet assignments and helps to correctly reject most of the implausible hypotheses. The tracker Acc+Ori is leveraging the IMU signals for the features and the reconstruction method of Section 3.5.2.4. The video information is used only to compare orientations between a video and the IMU signals. Except for the full VIT tracker, the variant Ori+Acc performs best among all other tracker variants. Its MOTA score is only about 20% worse than VIT. It achieves a very high IDF1 score of 76.7%, being about 16% worse than VIT. Note that Acc+Ori is independent of any artificial motion model or of the constancy assumptions on the appearance, as it just compares measurements. This shows the potential of the VIMPT setting.

By considering all features, which corresponds to our proposed VIT approach, we obtain the highest IDF1 score of 91.2%. In this case, the rejection of implausible hypotheses pairs based on acceleration is more meaningful.

Visual Heading Network accuracy. We evaluate the accuracy of the Visual Heading Network by computing the relative number of predicted heading vectors $\hat{\mathbf{n}}_\mathbf{d}$ that deviate no more than ϵ degrees from ground truth. The network is trained on the VIMPT2019 training sequence and tested on all other sequences of the dataset. According to Table 3.10, the network predicts orientations with high accuracy and is able to generalize to unseen images using the proposed perspective correction (Train[PC]

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a) **(b)**

Figure 3.15: Impact of different reconstruction methods when detections are missing.

Table 3.9: Tracking accuracy for different variants of our proposed tracker (VIT), evaluated on the VIMPT2019 test set.

Tracker	IDP[%] ↑	IDR[%] ↑	IDS ↓	MOTA[%] ↑	IDF1[%] ↑
VT	38.0	38.1	267	52.0	38.1
VT+Acc	44.9	45.1	256	65.0	45.0
Ori	48.1	49.3	1083	45.1	48.7
VT+Ori	77.0	75.8	146	58.9	76.4
Acc+Ori	76.5	77.0	322	71.6	76.7
VIT	**92.9**	**89.6**	**66**	**85.3**	**91.2**

and Test[PC]). In addition, the evaluations show that the perspective correction (PC) is crucial to obtain accurate results on the test set. Without the perspective correction (Train[w/o PC] and Test[w/o PC]), the neural network is unable to generalize the observed perspectives of the training data, thus being heavily prone to overfitting.

Since the orientation feature has shown to be very discriminative, the VHN is key to our proposed tracking approach.

Runtime. In general, solving binary quadratic problems such as the weighted graph labeling problem $\text{WGL}_{\text{NMS}}(\mathcal{G})$ is very challenging. However, in our experiments we observed very fast solutions. On an Intel i9 CPU with 8 cores and 3.60GHz, the runtime of the solver[15] for the entire dataset was 28 seconds. We attribute this to the IMU features being very discriminative, resulting in a very constrained optimization problem. Note that the complexity of $\text{WGL}_{\text{NMS}}(\mathcal{G})$ increases with the number of people to be tracked. If runtime becomes critical, an approximate solver (see Section 3.3 for a discussion and alternatives) can be applied to accelerate the computations.

---

[15] We used Gurobi in version 9.02.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 3.10: Training and test accuracy of the Visual Heading Network. We provide the relative number of heading errors within a threshold of $\epsilon \in \{30°, 45°\}$.

	$\leq 30° \uparrow$	$\leq 45° \uparrow$
Train[PC]	**88.8%**	**97.2%**
Train[w/o PC]	87.5%	96.7%
Test[PC]	**88.1%**	**96.2%**
Test[w/o PC]	70.4%	86.3%

Identification accuracy. While the ID metrics IDP, IDR, and IDF1 assess the consistency of assigning detections to person IDs, this does not necessarily mean that a person's trajectory is assigned the label of the corresponding IMU device of that person. To evaluate this, we utilize the matching established by the ID metrics (see Section 2.7.4). The assigned label of a computed trajectory is compared to the label corresponding to the IMU device of the person generating the matched ground truth trajectory. The VIT method produces 100% identification accuracy on the VIMPT dataset, *i.e.*, each computed trajectory is correctly assigned to the person generating the trajectory in terms of the IMU ID.

The proposed method thus simultaneously tracks and identifies IMU-equipped people recorded in a video.

Qualitative results. Tracking results are presented in Figure 3.9 and Figure 1.1. They show that despite heavy occlusions, motion blur, and similar appearance in the football sequences, VIT accurately tracks all persons, producing long-term consistent trajectories. In addition, the method is very robust to occlusions and can handle situations of people leaving and re-entering the scene. For instance in the park sequence depicted in the bottom row of Figure 3.9, the persons with assigned ID 4 and 7 are correctly re-identified after leaving the scene for about 10 seconds. As VIT performs person identification independent of the outward appearance, all depicted persons in Figure 3.9 are automatically and correctly assigned the same IDs in both shown sequences despite heavy changes in their visual appearances.

3.6 Conclusion

The holistic fusion of two complementary signals within an HO-MOT formulation allows overcoming weaknesses of individual signals, which is demonstrated in this chapter. For video-based MPT, two types of image features, namely head detections and people detections, are combined in a weighted graph labeling formulation. The underlying binary quadratic optimization problem is shown to be $\mathcal{NP}$-hard. To approximate an optimal solution, the Frank-Wolfe algorithm can be used. Yet, a simple application on the continuous relaxation is not sufficient to ensure good tracking results. The approximation quality is improved in three ways: (i) an algebraically exact and optimal step size is presented that is efficient to compute and (ii) a regularization is proposed that results in better discrete solutions. Finally, (iii) a novel hierarchical solving scheme

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

improves a feasible solution by correcting approximation errors. It is easy to integrate, fast to compute, and delivers solutions close to optimality. The resulting tracking method *Frank-Wolfe Tracker* (FWT) demonstrates the impact of these contributions.

Experiments show that the proposed solver modifications lead almost to a doubling of the MOTA metric as well as the objective value. Provided solutions are close to the optimum, having an approximation error of less than 1% in the experiments. At the same time, the proposed solver is considerably faster than an optimal BQP solver, which even after 35 times the calculation time of the approximate solver still has a worse solution. Even more, the experiment was performed on a reduced instance size using only people detections to enable the usage of the optimal solver.

Most importantly, conducted experiments show the advantage of incorporating two signals for MOT compared to the traditional tracking-by-detection approach. By using complementary signals, people detections that are not consistent with the heads can be recognized and removed, leading to a reduction of false positive detections by more than 50%. The overall tracking quality is significantly improved by about 15% in terms of MOTA. Consequently, FWT shows state-of-the-art results on two challenging tracking benchmarks compared to competing MOT methods at the time the work was published. In addition, the method won the MOT 2017 Tracking challenge at the CVPR 2017.

Despite the good performance, the FWT method is subject to some limitations. The experiments revealed an increased number of ID switches (IDS). We have identified the employed pairwise features, especially DeepMatching as the main error source. This can be tackled in future work with more advanced image features that better exploit the available image information, *e.g.*, by using a more global view.

Also, FWT needs an upper bound on the number of objects to be tracked. However, it can be viewed as an initialization that does not need to be exact, since the hierarchical solving scheme can compensate for a number that is chosen too small. If only detections shall be tracked without signal fusion, the next chapter of this thesis presents an HO-MOT method that bypasses the problem of estimating the number of objects in advance, using node-disjoint paths to represent trajectories.

Yet, some setups allow setting the upper bound exactly in advance. This is studied in the second application of signal fusion for MPT in this chapter. Here, video information is combined with measurements from body-worn IMUs for the purpose of multiple people tracking, which we term *Video Inertial Multiple People Tracking* (VIMPT). Conceptually, the setup enables accurate long-term tracking of multiple people even under dynamic motions and heavy occlusions. An interesting characteristic of VIMPT is that video-based trajectories of objects equipped with an IMU have to be assigned to the respective IMU devices. Hence, given that the correspondences between objects and IMU devices are known, a tracking solution automatically provides object identities within a video. To tackle the challenging VIMPT problem, the graph labeling formulation $\mathrm{WGL}_{\mathrm{NMS}}(\mathcal{G})$ is used to assign tracklets in a video to corresponding IMU devices such that the assignments are consistent with both the video information and the IMU signals.

The resulting tracking method *Video Inertial Tracker* (VIT) links IMU orientations with regressed orientations of persons in a video. For this purpose, a neural network with a novel perspective correction procedure is proposed that takes into account the

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

position of a person relative to the camera. The perspective correction significantly increases the accuracy of the orientation prediction. If an error tolerance of 45° is allowed, the improvement is about 12% compared to the neural network that ignores these perspective effects. In addition, VIT correlates accelerations measured from the IMU sensors with initial velocities, as measured in the corresponding video.

Besides the proposed tracking approach VIT, we released a VIMPT dataset, called VIMPT2019, which is used to evaluate the effectiveness of the VIMPT setting and to benchmark the proposed tracker. Conducted experiments on the recorded dataset show that the integration of orientation and acceleration information substantially improves tracking results, as the ID consistency (IDF1) increases by nearly 140% when IMU information is exploited. With a score of 91.2% IDF1, very reliable tracking results are achieved. The automated labeling of video trajectories (in terms of IMU devices) works without any error. Current state-of-the-art video-based trackers mainly rely on appearance information to establish a similarity measure between person detections. However, there are situations in which people wear similar or even identical apparel, and appearance is less informative. This observation was the main motivation to record VIMPT2019 and to propose an IMU enhanced tracking solution that is independent of person appearance and still able to track fast and dynamic motions. The experiments reveal that VIT performs significantly better than purely video-based tracking methods. In particular, the IDF1 score is improved by more than 100% compared to the best competing purely video-based MPT method on the challenging soccer sequences of the VIMPT2019 dataset.

Also, the combination of video and IMU data makes tracking less susceptible to impaired visual information. This is quantified by synthetically removing input detections and then recovering missing detections based on tracking results. Here, video-based linear interpolation is compared against the proposed fusion-based reconstruction method of VIT. The evaluations on the VIMPT2019 test set reveal that VIT is significantly more robust when both input signals are exploited. For example, removing only 20% of the input detections results already in a dramatic drop in tracking accuracy when using video-based linear interpolation. The MOTA value drops from 60% to -1% and the metric IDF1 drops from 78% to 43%. In contrast, VIT shows much better robustness: the IDF1 value drops from 91% to 81% and the MOTA value from 86% to 77%. The conducted experiments thus demonstrate the potential of the VIMPT setting and the proposed VIT method.

Yet, VIT shows some limitations with respect to practicability. Every object or person has to be equipped with an IMU which is impractical in certain situations. Also, the intrinsics and extrinsics of the camera have to be calibrated.

Overall, this chapter demonstrates that the fusion of object detections with additional signals is beneficial. Significant improvements were measured for video-based tracking by incorporating additional information retrieved from the images as well as for the fusion of two modalities. The proposed solver is essential to guarantee good tracking results. The presented VIT method integrates IMUs into an MPT framework in order to improve accuracy and to simultaneously obtain person assignments, which worked error-free.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

4 Lifted Disjoint Paths[1]

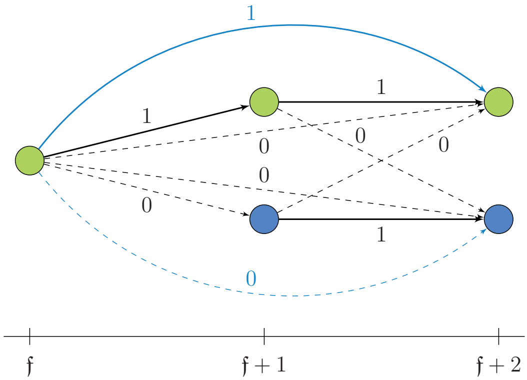


Figure 4.1: Illustration of connections between detections in three consecutive video frames together with a feasible labeling. We enhance the network flow model (black) with lifted edges (blue) representing whether two detections belong to the same track or not. Solid edges labeled by 1 represent active connections, dashed edges labeled by zero represent inactive connections between detections. Equally colored nodes represent detections of the same object.

This chapter presents an HO-MOT method that incorporates long-range temporal interactions and higher-order consistencies while avoiding suboptimal choices during optimization. To this end, an extension to the disjoint paths problem is proposed in which additional *lifted* edges are introduced to provide path connectivity priors. We call the resulting optimization problem the *lifted disjoint paths problem*. It is shown that this problem is $\mathcal{NP}$-hard by reduction from integer multicommodity flow and 3-SAT. To enable practical global optimization, several classes of linear inequalities are proposed that produce a high-quality LP-relaxation. Additionally, efficient cutting plane algorithms for separating the proposed linear inequalities are presented. The lifted disjoint paths problem is a natural model for multiple object tracking and allows an elegant mathematical formulation for long-range temporal interactions. Lifted edges

---

[1] This chapter contains text, images and results of previously published work [50].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

help to prevent ID switches and to re-identify persons. Previous methods either employ heuristics on similarly expressive models or use global optimization on simple models that do not incorporate long-range interactions. Compared with the method introduced in Chapter 3, no upper bound on the number of expected objects needs to be determined and the solver proposed in this chapter delivers provable global optimal solutions.

The proposed tracker achieves nearly optimal assignments with respect to given input detections. As a consequence, the method achieved significant improvements over the state of the art on all considered MOT datasets when the work was published [50].

4.1 Introduction

A prerequisite to obtain accurate MOT results is to employ a data association that is as robust as possible to detection errors (see Section 1.3.1) as well as to misleading weights caused by non-discriminative features (see Section 1.3.2). The approach of Chapter 3 tackles these challenges by extending the traditional tracking-by-detection paradigm to include additional signals holistically. This chapter addresses the problem from another perspective by trying to leverage the information of the provided detections as well as possible.

The proposed method builds upon the *Disjoint Paths* (DP) problem, *i.e.*, a special case of the network flow problem with flows constrained to be binary. DP applied to MOT [61] allows solving even very large data association instances to global optimality, see also Figure 1.5. Yet, tracking accuracies are limited as costs only indicate whether two detections directly follow each other in a track. To overcome the limited expressiveness of disjoint paths, we propose to augment it with *lifted* edges, which take into account long-range interactions, see Figure 4.1. We call the resulting problem the *lifted disjoint paths* (LDP) problem, which is presented in Section 4.3.

The LDP formulation has advantages from the modeling and optimization point of view. From the modeling standpoint, the lifted disjoint paths problem does not change the set of feasible solutions but adds more expressive power to it. For MOT, this means that the set of feasible solutions, which naturally represent trajectories of objects, is preserved. The additional lifted edges represent connectivity priors, see also Section 1.2.2. A lifted edge is active if and only if there is an active trajectory between its endpoints in the flow graph. For MOT, lifted edges take (dis-)similarity of object detection pairs represented by its endpoints into account. This allows to encourage or penalize an active path between the detections with possibly larger temporal distance. This helps to re-identify the same object and to prevent ID-switches between distinct objects within long trajectories.

From the optimization point of view, we show that LDP allows solving implicit HO-MOT problems via a global approach, in contrast to established approaches, which either use heuristics on complex models or global optimization on simpler models that do not exploit long-range interactions. To our knowledge, the presented LDP solver is the first global optimization approach that incorporates long-range interactions for MOT. This has several advantages: First, the optimization is not trapped in poor local optima or affected by initialization choices and is hence potentially more robust.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Second, improvements in the discriminative power of features used to compute costs for the lifted disjoint paths problem directly correlate to better tracking performance, since no errors are introduced by suboptimal choices during optimization.

To obtain a global optimal solver for LDP, several non-trivial classes of linear inequalities are studied that result in a high-quality relaxation. The proposed inequalities depend non-trivially on the constraint structure of the underlying disjoint paths problem, see Section 4.4. It is shown that the proposed polyhedral relaxation is tighter than naively applying known inequalities. These constraints are integrated into separation routines, see Section 4.5. At the same time, it is proven that LDP is $\mathcal{NP}$-hard in Section 4.6. Note also that the proposed lifted disjoint paths formulation is not inherently tied to MOT and can potentially be applied to further problems not related to MOT.

Finally, LDP is applied to MOT in Section 4.7. Spatio-temporal and appearance cues are combined in a neural network producing accurate pairwise costs for detections being up to 2 seconds apart. By incorporating long-range interactions using these long-term connections, the consistency (IDF1) of the result produced by the LDP method is substantially improved. Consequently, the method significantly outperformed state-of-the-art trackers on the MOT15/16/17 datasets at the time the work was published [50].

In summary, this chapter presents a method to improve the tracking accuracy by incorporating long-range temporal interactions, higher-order consistencies, and avoiding suboptimal decisions during the optimization. This is achieved by:

- Extending the disjoint paths problem by lifted edges to incorporate path connectivity priors, which we call the lifted disjoint paths problem. While the natural trajectory representation of disjoint paths is maintained, the proposed model is more expressive.
- Proposing accurate pairwise features that are stable across long temporal distances.
- The first global optimal solver for implicit HO-MOT taking long-range interactions into account. It is based on efficient cutting-plane algorithms and produces high-quality solutions. The tracker takes connections between detections being up to 60 frames apart into account.

4.2 Related Work

Disjoint paths problem. The disjoint paths problem can be solved with fast combinatorial solvers [238]. The shortest paths method for network flow specialized for the disjoint paths problem [108] performs extremely well in practice. For the case of the two disjoint paths problem, the specialized combinatorial algorithm by Suurballe's [239] can be used.

There exist several $\mathcal{NP}$-complete extensions to the disjoint paths problem. The shortest disjoint paths problem with multiple source-sink pairs [240] is $\mathcal{NP}$-complete, as is the more general integer multicommodity flow problem [241]. The special case of the disjoint paths problem with two distinct source/sink pairs, however, can be solved in polynomial time [242].

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Connectivity priors & lifted edges. For several combinatorial problems, special connectivity-inducing edges, which we will call lifted edges for our problem, have been introduced to improve expressiveness of the base problem, see also Figure 1.10.

In the Markov Random Field literature, special connectivity inducing edges were studied from a polyhedral point of view [243]. They were used in image analysis to indicate that two non-adjacent pixels belong to the same object, and hence they must be part of a contiguously labeled component of the underlying graph.

For multicut (a.k.a. correlation clustering), a classical graph decomposition problem, lifted edges have been introduced by Keuper *et al.* [244] to model connectivity priors and analyzed in terms of its polyhedral properties [245]. A lifted edge expresses affinity of two nodes to be in the same/different connected component of the graph partition. Lifted multicut has been used for image and mesh segmentation [244], connectomics [246], and cell tracking [40]. A combination of the lifted multicut problem and Markov Random Fields has been proposed [51] with applications in instance-separating semantic segmentation [247].

Yet, for the above problems, global optimization has only been reported for small instances.

Disjoint paths for MOT. The data association step of MOT has been approached using the disjoint paths setup [61, 67], since disjoint paths through a graph naturally model trajectories of multiple objects (see Figure 1.5 and Section 1.2.2). Extensions of the plain disjoint paths problem that disallow certain pairs of detections to occur simultaneously have been used to fuse different object detectors [103] and for multi-camera MOT [110, 248]. The drawback of these approaches is that they cannot integrate long-term information, in contrast to our proposed formulation.

Other combinatorial approaches to MOT. The minimum cost arborescence problem, an extension of minimum spanning tree to directed graphs, has been used [106] for MOT. Several works [54, 63, 77, 249–251] employ the multicut problem for MOT. Additionally, lifted edges have been used [64, 252] to better model long-range temporal interactions. The maximum clique problem [52, 70], which corresponds to multicut with complete graphs, has been applied for MOT, see also Figure 1.18. Maximum independent set, which corresponds to maximum clique on the complement graph, has been used by Brendel *et al.* [53] for MOT. The multigraph-matching problem, a generalization of the graph matching problem, has been applied to MOT by Hu *et al.* [253]. Consistency of individually matched detections is ensured by cycle-consistency constraints coming from multi-graph matching. Section 3 reformulates tracking multiple objects with long temporal interactions as a binary quadratic program and solves it with a specialized non-convex Frank-Wolfe method. Common to the above state-of-the-art trackers is that they either employ heuristic solvers or are limited in the integration of long-range information, in contrast to the method of this chapter.

Contribution w.r.t. existing combinatorial approaches. It is widely acknowledged that one crucial ingredient for obtaining high-quality MOT results is to incorporate

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

long-term temporal information to re-identify detections and prevent ID-switches. However, from a theoretical perspective, we believe that long-range information has not yet been incorporated satisfactorily in optimization formulations for the data association step in MOT.

In comparison to lifted multicut for MOT, we argue that from the modeling point of view, network flow has advantages. In multicut, clusters can be arbitrary, while in MOT, tracks are clusters that may not contain multiple detection hypotheses of distinct objects at the same point in time. This exclusion constraint must be enforced in multicut explicitly via soft constraints, while the disjoint paths substructure automatically takes care of it. On the other hand, the lifted multicut approach [64] has used the possibility to cluster multiple detections in one time frame. This directly incorporates non-maxima suppression in the optimization, which however increases computational complexity.

From a mathematical perspective, naively using polyhedral results from multicut is also not satisfactory. Specifically, one could naively obtain a polyhedral relaxation for the lifted disjoint paths problem by reusing the known polyhedral structure of lifted multicut [245] and additionally adding network flow constraints for the disjoint paths substructure. However, this would give a suboptimal polyhedral relaxation. We show in Section 4.4 that the underlying structure of the disjoint paths problem can be used to derive new and tighter constraints for lifted edges. This enables us to use a global optimization approach for MOT. To our knowledge, our work is the first to combine global optimization with long-range interactions for MOT.

In comparison to works that propose non-convex algorithms or other heuristics for incorporating long-range temporal edges [52, 70, 253], but also compared to the proposed solver of Chapter 3, the approach of this chapter yields a more principled approach and globally optimal optimization solutions via LP-based branch-and-bound algorithms.

4.3 Problem Formulation

Below we recapitulate the disjoint paths problem and extend it by defining lifted edges. We discuss how the lifted disjoint paths problem can naturally model MOT.

Flow network and lifted graph. Consider two directed acyclic graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and $\breve{\mathcal{G}} = (\breve{\mathcal{V}}, \breve{\mathcal{E}})$, where $\breve{\mathcal{V}} = \mathcal{V} \backslash \{s, t\}$. The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represents the *flow network* and we denote by $\breve{\mathcal{G}}$ the lifted graph. The two special nodes s and t of $\mathcal{G}$ denote the source and sink node, respectively. We further assume that every node in $\mathcal{V}$ is reachable from s, and t can be reached from it.

We define the set of paths starting at v and ending in w as

$$vw\text{-paths}(\mathcal{G}) = \left\{ (v_1 v_2, \ldots, v_{l-1} v_l) : \begin{array}{c} v_i v_{i+1} \in \mathcal{E}, \\ v_1 = v, v_l = w \end{array} \right\}. \tag{4.1}$$

For a vw-path P we denote its edge set as $P_{\mathcal{E}}$ and its node set as $P_{\mathcal{V}}$.

The flow variables in $\mathcal{G}$ are denoted by $y \in \{0,1\}^{\mathcal{E}}$ for edges and $x \in \{0,1\}^{\mathcal{V}}$ for nodes. Allowing only 0/1 values of vertex variables reflects the requirement of vertex disjoint

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

paths. Variables on the lifted edges $\breve{\mathcal{E}}$ are denoted by $\breve{y} \in \{0,1\}^{\breve{\mathcal{E}}}$. Here, $\breve{y}_{vw} = 1$ means that nodes v and w are connected via the flow y in $\mathcal{G}$. Formally,

$$\breve{y}_{vw} = 1 \Leftrightarrow \exists P \in vw\text{-paths}(\mathcal{G}) \text{ s.t. } \forall ij \in P_{\mathcal{E}} : y_{ij} = 1\,. \tag{4.2}$$

Optimization problem. Given edge costs $\mathbf{q} \in \mathbb{R}^{\mathcal{E}}$, node costs $\mathbf{c} \in \mathbb{R}^{\mathcal{V}}$ in flow network $\mathcal{G}$, and edge costs $\breve{\mathbf{q}} \in \mathbb{R}^{\breve{\mathcal{E}}}$ for the lifted graph $\breve{\mathcal{G}}$, we define the (minimum cost) lifted disjoint paths problem as

$$\begin{aligned} \min_{\substack{\mathbf{y}\in\{0,1\}^{\mathcal{E}},\breve{\mathbf{y}}\in\{0,1\}^{\breve{\mathcal{E}}},\\ \mathbf{x}\in\{0,1\}^{\mathcal{V}}}} \quad & \langle \mathbf{q}, \mathbf{y}\rangle + \langle \breve{\mathbf{q}}, \breve{\mathbf{y}}\rangle + \langle \mathbf{c}, \mathbf{x}\rangle \\ \text{s.t.} \quad & \mathbf{y} \text{ node-disjoint } s,t\text{-flow in } \mathcal{G}\,, \\ & \mathbf{x} \text{ flow through nodes of } \mathcal{G}\,, \\ & \mathbf{y}, \breve{\mathbf{y}} \text{ feasible according to (4.2)}\,. \end{aligned} \tag{4.3}$$

In Section 4.4, we present a BLP formulation of (4.3) by proposing several linear inequalities that lead to a high-quality linear relaxation.

Graph construction for multiple object tracking. We argue that the lifted disjoint paths problem is an appropriate way of modeling the data association problem for MOT. In MOT, an unknown number of objects needs to be tracked across a video sequence. This problem can be naturally formalized by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where its node set $\mathcal{V}$ represents either object detections or tracklets of objects. If $\mathcal{V}$ represents object detections, we can express it as follows: $\mathcal{V} = \{s\} \cup \mathcal{V}_1 \cup \ldots \cup \mathcal{V}_{n_{\mathrm{R}}} \cup \{t\}$, where n_{R} is the number of frames and $\mathcal{V}_{\mathfrak{f}}$ denotes the object detections in frame $\mathfrak{f}$. We introduce edges between adjacent time frames. An active flow on such an edge denotes correspondences of the same object. We also introduce skip edges between time frames that are farther apart. An active flow on a skip edge also denotes correspondences between the same object that, in contrast, may have been occluded or not detected in intermediate time frames. This classical network flow formulation has been commonly used for MOT [61].

On top of the underlying flow formulation for MOT, we usually want to express that two detections belong to the same object connected by a possibly longer track with multiple detections in between. For that purpose, lifted edges with negative costs can be used. We say in such a case that an active lifted edge re-identifies two detections [64]. If two detections with larger temporal distance should not be part of the same track, a positive valued lifted edge can be used. In this case, the lifted edge is used to prevent ID-switches.

4.4 Constraints

Below, we will first introduce constraints that give a binary linear program (BLP) of the lifted disjoint paths problem (4.3). The corresponding linear programming (LP) relaxation can be strengthened by additional constraints that we present subsequently.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Many constraints considered below will rely on whether a node w is reachable from another node v in the flow network. We define to this end the *reachability relation* $\mathcal{R} \subset \mathcal{V}^{(2)}$ via

$$vw \in \mathcal{R} \Leftrightarrow vw\text{-paths}(\mathcal{G}) \neq \emptyset\,. \tag{4.4}$$

In the special case of $v = w$, we also allow empty paths, which means $\forall v \in \mathcal{V} : vv \in \mathcal{R}$. This makes relation $\mathcal{R}$ reflexive.

Flow conservation constraints. The flow variables y obey, as in classical network flow problems [107], the flow conservation constraints

$$\forall v \in \mathcal{V} \setminus \{s,t\} : \sum_{u: uv \in \mathcal{E}} y_{uv} = \sum_{w: vw \in \mathcal{E}} y_{vw} = x_v\,. \tag{4.5}$$

Constraining lifted edges. All the following constraints restrict values of lifted edge variables $\breve{y}_{vw}$ in order to ensure that they satisfy (4.2). Despite their sometimes complex form, they always obey the two basic principles:

- If there is flow in $\mathcal{G}$ going from vertex v to vertex w, then $\breve{y}_{vw} = 1$. The constraints of this form are (4.9) and (4.12).
- If there is a vw-cut in $\mathcal{G}$ with all edges labeled by zero (*i.e.*, no flow passes through this cut), then $\breve{y}_{vw} = 0$. We will mainly look at cuts that are induced by paths, *i.e.*, edges that separate a path from the rest of the graph. The paths of interest will either originate at v or end at w. The constraints of this form are (4.6), (4.7), (4.8), (4.13), and (4.14), see below.

Single node cut inequalities. Given a lifted edge $vw \in \breve{\mathcal{E}}$, if there is no flow going from vertex v which can potentially go to vertex w, then $\breve{y}_{vw} = 0$. Formally,

$$\breve{y}_{vw} \leq \sum_{\substack{u:\ vu \in \mathcal{E},\\ uw \in \mathcal{R}}} y_{vu}\,. \tag{4.6}$$

Similarly, if there is no flow going to w that can originate from vertex v, then $\breve{y}_{vw} = 0$. Formally,

$$\breve{y}_{vw} \leq \sum_{\substack{u: uw \in \mathcal{E},\\ vu \in \mathcal{R}}} y_{uw}\,. \tag{4.7}$$

The number of constraints of the above type (4.5) is linear in the number of vertices, while (4.6) and (4.7) are linear in the number of lifted edges. Hence we add them into our initial constraint set during optimization.

Path-induced cut inequalities. The path-induced cut inequalities generalize the single node cut inequalities (4.6) and (4.7) by allowing cuts induced by paths.

Let a lifted edge $vw \in \breve{\mathcal{E}}$, a node u from which w is reachable, and a vu-path P be given. Consider the cut given by edges ik with $i \in P_{\mathcal{V}}$ and $k \notin P_{\mathcal{V}}$ but such that w is

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

reachable from k. If the flow does not take any edge of this cut, then $\breve{y}_{vw} = 0$. Formally,

$$\begin{aligned} &\forall vw \in \breve{\mathcal{E}}\ \forall P \in vu\text{-paths}(\mathcal{G}) \text{ s.t. } uw \in \mathcal{R} \wedge u \neq w : \\ &\breve{y}_{vw} \leq \sum_{i \in P_{\mathcal{V}}} \sum_{\substack{k \notin P_{\mathcal{V}},\\ kw \in \mathcal{R}}} y_{ik} \,. \end{aligned} \tag{4.8}$$

Path inequalities. For lifted edge $\breve{y}_{vw}$, it holds that if there is a flow in $\mathcal{G}$ going from v to w along a path P, then $\breve{y}_{vw} = 1$. This constraint can be expressed by the following set of inequalities:

$$\begin{aligned} &\forall vw \in \breve{\mathcal{E}}\ \forall P \in vw\text{-paths}(\mathcal{G}) : \\ &\breve{y}_{vw} \geq \sum_{\substack{vj \in \mathcal{E},\\ j \in P_{\mathcal{V}}}} y_{vj} - \sum_{i \in P_{\mathcal{V}} \setminus \{v,w\}} \sum_{k \notin P_{\mathcal{V}}} y_{ik} \,. \end{aligned} \tag{4.9}$$

Here the first sum expresses the flow going from v to any vertex of path P. The second sum is the flow leaving path vertices $P_{\mathcal{V}}$ before reaching w. In other words, if flow does not leave $P_{\mathcal{V}}$, edge $\breve{y}_{vw}$ must be active.

Remark 4.1. Inequality (4.9) implicitly enforces $\breve{y}_{vw}$ to be active if any vw-path $\tilde{P}$ with $\tilde{P}_{\mathcal{V}} \subset P_{\mathcal{V}}$ is active and no flow goes through a skipped node $u \in P_{\mathcal{V}} \setminus \tilde{P}_{\mathcal{V}}$.

Lifted Disjoint Paths as a BLP. Using the cut inequalities (4.6), (4.8), and the path inequalities (4.9), Theorem 4.2 proves that the lifted disjoint paths problem is a BLP. The remainder of this section then presents additional inequalities such that the feasibility set of the lifted disjoint paths problem remains unchanged while the corresponding linear relaxations become tighter.

Theorem 4.2. *The lifted disjoint paths problem* (4.3) *is a BLP.*

Proof. We show that we can replace (4.2) by the linear constraints (4.6), (4.8), and (4.9) without changing the feasibility set. Let us assume that (4.6), (4.8), (4.9), and the flow conservation (4.5) hold.

- Let us prove $\breve{y}_{vw} = 1 \Longrightarrow \exists P \in vw\text{-paths}(\mathcal{G})$ s.t. $\forall ij \in P_{\mathcal{E}} : y_{ij} = 1$.
  According to the single node cut constraints (4.6), a node $u \in \mathcal{V}$ exists with $vu \in \mathcal{E}$, $y_{vu} = 1$, and $uw \in \mathcal{R}$.

  If $u = w$ holds, $P = (vu)$ satisfies the statement to be shown.

  Now assume $u \neq w$. Let $u_0 = v$, $u_1 = u$, and $\tilde{P} = (u_0u_1, \ldots, u_{n-1}u_n) \in vu_n$-paths$(\mathcal{G})$ be the path of maximal length such that (i) $y_{u_{i-1}u_i} = 1$ for all $i \in [n]$, (ii) $u_n \neq w$, and (iii) $u_n w \in \mathcal{R}$. Note that there exists a unique path $\tilde{P}$ with the desired property by the flow conservation constraints (4.5) and the single node cut constraints (4.6). For each $i \in [n-1]_0$, the flow conservation constraints (4.5) imply

$$0 = \sum_{\substack{k \notin \tilde{P}_{\mathcal{V}},\\ kw \in \mathcal{R}}} y_{u_i k} \,. \tag{4.10}$$

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

The path induced cut inequality constraints (4.8) thus ensure that

$$\breve{y}_{vw} = 1 \leq \sum_{\substack{k \notin \tilde{P}_{\mathcal{V}}, \\ kw \in \mathcal{R}}} y_{u_n k} \,. \tag{4.11}$$

It follows that there exists a node $k \in \mathcal{V} \setminus \tilde{P}_{\mathcal{V}}$ with $kw \in \mathcal{R}$ and $y_{u_n k} = 1$. Since $\tilde{P}$ was chosen with maximal length, $k = w$ must hold. The required path properties are thus satisfied by $P = (u_0 u_1, \ldots, u_{n-1} u_n, u_n w) \in vw\text{-paths}(\mathcal{G})$.

- Let us prove $\exists P \in vw\text{-paths}(\mathcal{G})$ s.t. $\forall ij \in P_{\mathcal{E}} : y_{ij} = 1 \implies \breve{y}_{vw} = 1$.
  Let $P \in vw\text{-paths}(\mathcal{G})$ be a path with $y_{ij} = 1$ for all $ij \in P_{\mathcal{E}}$. By the flow conservation constraints (4.5), $\exists! j \in P_V : y_{vj} = 1$. By assumption, $\forall i \in P_{\mathcal{V}} \setminus \{v, w\} : \exists j \in P_V : y_{ij} = 1$. Consequently, the flow conservation constraints (4.5) imply $\forall i \in P_{\mathcal{V}} \setminus \{v, w\} : \forall k \in \mathcal{V} \setminus P_V : y_{ik} = 0$. The assertion thus follows by the path inequalities (4.9).

□

Lifted inequalities. The path inequalities (4.9) and the path-induced cut inequalities (4.8) only consider base edges on their right-hand sides. We can generalize both (4.9) and (4.8) by including lifted edges in the paths as well. Conceptually, using lifted edges allows representing all possible paths between their endpoints, which enables to formulate tighter inequalities, see Propositions 4.3 and 4.4.

To that end, consider the multigraph $\mathcal{G} \cup \breve{\mathcal{G}} := (\mathcal{V}, \mathcal{E} \cup \breve{\mathcal{E}})$. For any edge $ij \in \mathcal{E} \cap \breve{\mathcal{E}}$, we always distinguish whether $ij \in \mathcal{E}$ or $ij \in \breve{\mathcal{E}}$. For $P \in vw\text{-paths}(\mathcal{G} \cup \breve{\mathcal{G}})$, we denote by $P_{\mathcal{E}}$ and $P_{\breve{\mathcal{E}}}$ edges of the path P in $\mathcal{E}$ and $\breve{\mathcal{E}}$ respectively. We require $P_{\mathcal{E}} \cap P_{\breve{\mathcal{E}}} = \emptyset$.

Lifted path inequalities. We generalize the path inequalities (4.9). Now the vw-path P may contain both edges in $\mathcal{E}$ and $\breve{\mathcal{E}}$. Whenever a lifted edge $\breve{y}_{ij}$ in the third sum in (4.12) is one, two cases can occur: (i) Flow goes out of P (uses vertices not in $P_{\mathcal{V}}$) but reenters it again later. Then a base edge variable y_{ik} will be one in the second sum in (4.12) and the values of $\breve{y}_{ij}$ and y_{ik} cancel out. (ii) A base edge $ij \in \mathcal{E} \cap \breve{\mathcal{E}}$ parallel to the lifted edge is active. Then the variable y_{ij} in the fourth sum in (4.12) cancels out $\breve{y}_{ij}$. The lifted path inequality becomes

$$\begin{aligned} &\forall vw \in \breve{\mathcal{E}} \ \forall P \in vw\text{-paths}(\mathcal{G} \cup \breve{\mathcal{G}}) : \\ &\quad \breve{y}_{vw} \geq \sum_{j \in P_{\mathcal{V}}} y_{vj} - \sum_{i \in P_{\mathcal{V}} \setminus \{v,w\}} \sum_{k \notin P_{\mathcal{V}}} y_{ik} \\ &\qquad\qquad + \sum_{ij \in P_{\breve{\mathcal{E}}}} \breve{y}_{ij} - \sum_{ij \in P_{\breve{\mathcal{E}}} \cap \mathcal{E}} y_{ij} \,. \end{aligned} \tag{4.12}$$

Whenever the path in (4.12) consists only of base edges $P_{\mathcal{E}}$, the resulting inequality becomes a path inequality (4.9).

Proposition 4.3. *The lifted path inequalities* (4.12) *provide a strictly better relaxation than the path inequalities* (4.9).

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Proof. Let us define the following sets

$$S_B = \{(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}} \mid (\mathbf{y}, \breve{\mathbf{y}}) \text{ satisfy (4.9)}\}\,,$$
$$S_L = \{(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}} \mid (\mathbf{y}, \breve{\mathbf{y}}) \text{ satisfy (4.12)}\}\,.$$

- Let us prove that $S_L \subset S_B$:
  Note that every path $P \in vw$-paths$(\mathcal{G})$ belongs to the set of vw-paths$(\mathcal{G} \cup \breve{\mathcal{G}})$ too. It just holds that $P_{\breve{\mathcal{E}}} = \emptyset$. Let us rewrite the right-hand side of the inequality from (4.12) for such $P \in vw$-path$(\mathcal{G} \cup \breve{\mathcal{G}})$ where $P_{\breve{\mathcal{E}}} = \emptyset$.

$$\begin{aligned} \breve{y}_{vw} &\geq \sum_{\substack{vj \in \mathcal{E},\\ j \in P_{\mathcal{V}}}} y_{vj} - \sum_{i \in P_{\mathcal{V}} \setminus \{v,w\}} \sum_{k \notin P_{\mathcal{V}}} y_{ik} \\ &\quad + \sum_{ij \in P_{\breve{\mathcal{E}}}} \breve{y}_{ij} - \sum_{ij \in P_{\breve{\mathcal{E}}} \cap \mathcal{E}} y_{ij} \\ &= \sum_{\substack{vj \in \mathcal{E},\\ j \in P_{\mathcal{V}}}} y_{vj} - \sum_{i \in P_{\mathcal{V}} \setminus \{v,w\}} \sum_{k \notin P_{\mathcal{V}}} y_{ik}\,. \end{aligned}$$

  This is exactly the right hand side of (4.9). Therefore, any pair of real vectors $(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}}$ that satisfies (4.12) must satisfy (4.9) as well.

- Let us prove that $S_L \subsetneq S_B$:
  We prove that there exists $(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}}$ such that $(\mathbf{y}, \breve{\mathbf{y}})$ satisfies (4.9) and does not satisfy (4.12). See the graph in Figure 4.2. There are four possible paths from v to w in $\mathcal{G}$. If we use Constraints (4.9), all the paths give us the same lower bound on $\breve{y}_{vw}$

$$\breve{y}_{vw} \geq 1 - 0.5 - 0.5 = 0\,.$$

  If we use Constraints (4.12) with path $P = (vv_1, v_1v_4, v_4w)$ where $P_{\breve{\mathcal{E}}} = \{v_1v_4, v_4w\}$, we obtain

$$\breve{y}_{vw} \geq 1 - 0.5 - 0.5 - 0.5 - 0.5 + 1 + 1 = 1\,.$$

□

Lifted path-induced cut inequalities. We generalize the path-induced cut inequalities (4.8). Let a lifted edge $vw \in \breve{\mathcal{E}}$ and a vu-path P in $\mathcal{G} \cup \breve{\mathcal{G}}$ be given. In contrast to the basic version (4.8), a lifted edge $ij \in P_{\breve{\mathcal{E}}}$ can be taken. This can occur in two cases: Either the flow leaves $P_{\mathcal{V}}$ via a base edge ik, $k \notin P_{\mathcal{V}}$ or a base edge $ij \in \mathcal{E} \cap \breve{\mathcal{E}}$ parallel to the lifted edge is taken. Both cases are accounted for by terms in the first and the third sum in (4.13) below.

$$\begin{gathered} \forall vw \in \breve{\mathcal{E}}\ \forall P \in vu\text{-paths}(\mathcal{G} \cup \breve{\mathcal{G}}) \text{ s.t. } \quad uw \in \mathcal{R} \wedge u \neq w : \\ \breve{y}_{vw} \leq \sum_{i \in P_{\mathcal{V}}} \sum_{\substack{k \notin P_{\mathcal{V}},\\ kw \in \mathcal{R}}} y_{ik} - \sum_{ij \in P_{\breve{\mathcal{E}}}} \breve{y}_{ij} + \sum_{ij \in P_{\breve{\mathcal{E}}} \cap \mathcal{E}} y_{ij}\,. \end{gathered} \tag{4.13}$$

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

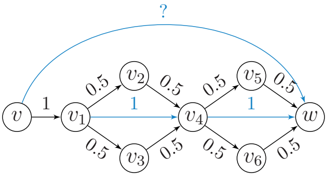


Figure 4.2: Exemplary case where the path inequalities (4.9) give a trivial lower bound on the lifted edge $\breve{y}_{vw}$. The lifted path inequality (4.12) gives the correct lower bound. Example for Proposition 4.3.

Assume that the last node u of path P is connected via a lifted edge with w. Then we can strengthen (4.13) by replacing the sum of base edges outgoing from u by $\breve{y}_{uw}$.

$$\begin{aligned} &\forall vw \in \breve{\mathcal{E}}\ \forall P \in vu\text{-paths}\ (\mathcal{G} \cup \breve{\mathcal{G}})\ \text{ s.t. } uw \in \breve{\mathcal{E}} : \\ &\breve{y}_{vw} \leq \sum_{i \in P_{\mathcal{V}} \setminus u} \sum_{\substack{k \notin P_{\mathcal{V}}, \\ kw \in \mathcal{R}}} y_{ik} - \sum_{ij \in P_{\breve{\mathcal{E}}}} \breve{y}_{ij} \\ &\qquad + \sum_{ij \in P_{\breve{\mathcal{E}}} \cap \mathcal{E}} y_{ij} + \breve{y}_{uw}\,. \end{aligned} \tag{4.14}$$

Proposition 4.4. *The lifted path-induced cut inequalities* (4.13) *define a strictly tighter relaxation than the path-induced cut inequalities* (4.8).

Furthermore, the lifted path-induced cut inequalities (4.13) *and* (4.14) *define a strictly better relaxation than* (4.13) *alone.*

Proof. Let us define the following sets

$$\begin{aligned} S_B &= \{(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}} \mid (\mathbf{y}, \breve{\mathbf{y}}) \text{ satisfy } (4.8)\}\,, \\ S_{L1} &= \{(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}} \mid (\mathbf{y}, \breve{\mathbf{y}}) \text{ satisfy } (4.13)\}\,, \\ S_{L2} &= \{(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}} \mid (\mathbf{y}, \breve{\mathbf{y}}) \text{ satisfy } (4.14)\}\,. \end{aligned}$$

- First, we prove $S_{L1} \subset S_B$:
  We use the same argument as in the proof of Proposition 4.3. Every path $P \in vw\text{-paths}(\mathcal{G})$ belongs to the set of vw-paths$(\mathcal{G} \cup \breve{\mathcal{G}})$ and it holds that $P_{\breve{\mathcal{E}}} = \emptyset$. Let us rewrite the right hands side of the inequality from (4.13) for such $P \in vw\text{-path}(\mathcal{G} \cup \breve{\mathcal{G}})$ where $P_{\breve{\mathcal{E}}} = \emptyset$.

$$\begin{aligned} \breve{y}_{vw} &\leq \sum_{i \in P_{\mathcal{V}}} \sum_{\substack{k \notin P_{\mathcal{V}} \\ kw \in \mathcal{R}}} y_{ik} - \sum_{ij \in P_{\breve{\mathcal{E}}}} \breve{y}_{ij} + \sum_{ij \in P_{\breve{\mathcal{E}}} \cap \mathcal{E}} y_{ij} \\ &= \sum_{i \in P_{\mathcal{V}}} \sum_{\substack{k \notin P_{\mathcal{V}} \\ kw \in \mathcal{R}}} y_{ik}\,. \end{aligned}$$

  Which is exactly the right hand side of (4.8). Therefore, any pair of real vectors $(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}}$ that satisfies (4.13) must satisfy (4.8).

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

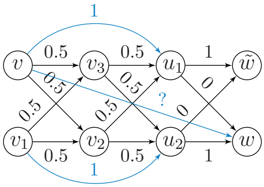


Figure 4.3: Exemplary case where the path-induced cut inequalities (4.8) fail to give non-trivial upper bounds for lifted edge $\breve{y}_{vw}$. The lifted path-induced cut-inequalities (4.13) give the correct upper bound in this case. Example for Proposition 4.4.

- Let us prove $S_{L1} \subsetneq S_B$:
  We prove that there exists $(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}}$ such that $(\mathbf{y}, \breve{\mathbf{y}})$ satisfies (4.8) and does not satisfy (4.13).

  See the example in Figure 4.3. There are four possible paths in $\mathcal{G}$ from v to either u_1 or u_2. They are $P_1 = (vv_3, v_3u_1)$, $P_2 = (vv_2, v_2u_1)$, $P_3 = (vv_3, v_3u_2)$, $P_4 = (vv_2, v_2u_2)$. Using (4.13), all of them give us the same threshold on $\breve{y}_{vw}$:

  $$\breve{y}_{vw} \leq 0.5 + 0.5 + 0 = 1\,.$$

  If we use Constraint (4.13) with path $P = (vu_1)$, we obtain the following threshold:

  $$\breve{y}_{vw} \leq 0.5 + 0.5 + 0 - 1 = 0\,.$$

- Let us prove that $S_{L1} \cap S_{L2} \subsetneq S_{L1}$
  It holds trivially that $S_{L1} \cap S_{L2} \subset S_{L1}$. Let us prove that there exists $(\mathbf{y}, \breve{\mathbf{y}}) \in [0,1]^{\mathcal{E}} \times [0,1]^{\breve{\mathcal{E}}}$ such that $(\mathbf{y}, \breve{\mathbf{y}}) \in S_{L1}$ and $(\mathbf{y}, \breve{\mathbf{y}}) \notin S_{L1} \cap S_{L2}$.

  See the example graph in Figure 4.4. Similarly as in Figure 4.3, there are four possible paths from v to either u_1 or u_2 in $\mathcal{G}$. There are no active lifted edges that would enable us to obtain a better upper bound on $\breve{y}_{vw}$ using (4.13) than the following:

  $$\breve{y}_{vw} \leq 1\,.$$

  However, if we use Constraints (4.14) with path $P = (vv_3)$ and $\breve{y}_{v_3w} = 0$, we obtain

  $$\breve{y}_{vw} \leq 0\,.$$

□

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

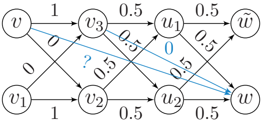


Figure 4.4: Exemplary failure case for the lifted path-induced cut inequalities (4.13). The lifted path-induced cut inequalities (4.14) give the correct upper bound for lifted edge $\breve{y}_{vw}$. Example for Proposition 4.4.

Symmetric cut inequalities. Inequalities (4.7) provide a symmetric counterpart to inequalities (4.6). We can also formulate symmetric counterparts to inequalities (4.8), (4.13), and (4.14) by swapping the role of v and w. All constraints (4.8), (4.13), and (4.14) concentrate on paths originating in v. The symmetric inequalities are obtained by studying all paths ending in w. Relations analogous to those described in Proposition 4.4 hold for the symmetric counterparts as well. The symmetric inequalities also strengthen the relaxation strictly.

Inequalities symmetric to (4.8):

$$\begin{aligned} &\forall vw \in \breve{\mathcal{E}}\ \forall P \in uw\text{-paths}(\mathcal{G}) \text{ s.t. } \quad vu \in \mathcal{R} \wedge u \neq v : \\ &\breve{y}_{vw} \leq \sum_{i \in P_{\mathcal{V}}} \sum_{\substack{k \notin P_{\mathcal{V}}, \\ vk \in \mathcal{R}}} y_{ki} \,. \end{aligned} \tag{4.15}$$

Inequalities symmetric to (4.13):

$$\begin{aligned} &\forall vw \in \breve{\mathcal{E}}\ \forall P \in uw\text{-paths}(\mathcal{G} \cup \breve{\mathcal{G}}) \text{ s.t. } \quad vu \in \mathcal{R} \wedge u \neq v : \\ &\breve{y}_{vw} \leq \sum_{i \in P_{\mathcal{V}}} \sum_{\substack{k \notin P_{\mathcal{V}}, \\ vk \in \mathcal{R}}} y_{ki} - \sum_{ij \in P_{\breve{\mathcal{E}}}} \breve{y}_{ij} \\ &\qquad + \sum_{ij \in P_{\breve{\mathcal{E}}} \cap \mathcal{E}} y_{ij} \,. \end{aligned} \tag{4.16}$$

Inequalities symmetric to (4.14):

$$\begin{aligned} &\forall vw \in \breve{\mathcal{E}}\ \forall P \in uw\text{-paths}\ (\mathcal{G} \cup \breve{\mathcal{G}}) \ \text{ s.t. } vu \in \breve{\mathcal{E}} : \\ &\breve{y}_{vw} \leq \sum_{i \in P_{\mathcal{V}} \setminus u} \sum_{\substack{k \notin P_{\mathcal{V}}, \\ vk \in \mathcal{R}}} y_{ki} - \sum_{ij \in P_{\breve{\mathcal{E}}}} \breve{y}_{ij} \\ &\qquad + \sum_{ij \in P_{\breve{\mathcal{E}}} \cap \mathcal{E}} y_{ij} + \breve{y}_{vu} \,. \end{aligned} \tag{4.17}$$

Proposition 4.5. *The lifted path-induced cut inequalities* (4.16) *define a strictly tighter relaxation than the path-induced cut inequalities* (4.15).
The lifted path-induced cut inequalities (4.16) *and* (4.17) *define a strictly better relaxation than* (4.16) *alone.*

Proof. Analogical to the proof of Proposition 4.4. See Figure 4.6 for example analogical to the one in Figure 4.3 and Figure 4.7 for example analogical to the one in Figure 4.4. □

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Proposition 4.6.

1. *The path-induced cut inequalities* (4.8) *together with their symmetric counterpart* (4.15) *define a strictly tighter relaxation than inequalities* (4.8) *alone.*
2. *The path-induced cut inequalities* (4.13) *together with their symmetric counterpart* (4.16) *define a strictly tighter relaxation than inequalities* (4.13) *alone.*
3. *Using path-induced cut inequalities* (4.17) *together with* (4.13), (4.14), *and* (4.16) *strictly improves the relaxation.*

Proof.

1. See the example in Figure 4.5.
   Upper bound on $\breve{y}_{vw}$ by (4.8): $\breve{y}_{vw} \leq 0.5 + 0.5 = 1$.
   Upper bound on $\breve{y}_{vw}$ by (4.15): $\breve{y}_{vw} \leq 0$.
2. See the example in Figure 4.6.
   Upper bound on $\breve{y}_{vw}$ by (4.13): $\breve{y}_{vw} \leq 0.5 + 0.5 = 1$.
   Upper bound on $\breve{y}_{vw}$ by (4.16) using path $P = (u_2 w)$: $\breve{y}_{vw} \leq 0 + 0.5 + 0.5 - 1 = 0$.
3. See the example in Figure 4.7.
   Upper bounds on $\breve{y}_{vw}$ by (4.13), (4.14), (4.16): $\breve{y}_{vw} \leq 1$.
   Upper bound on $\breve{y}_{vw}$ by (4.17) using path $P = (uw)$ and $\breve{y}_{vu} = 0$: $\breve{y}_{vw} \leq 0$.

□

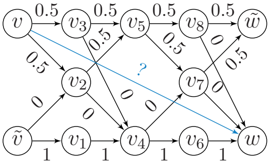


Figure 4.5: The best upper bound on $\breve{y}_{vw}$ is provided by inequalities (4.15). Example for Proposition 4.6.

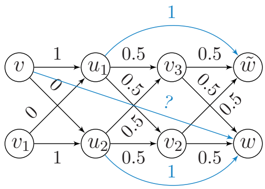


Figure 4.6: The best upper bound on $\breve{y}_{vw}$ is provided by inequalities (4.16). Example for Proposition 4.5 and Proposition 4.6.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

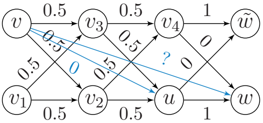


Figure 4.7: The best upper bound on $\breve{y}_{vw}$ is provided by inequalities (4.17). Example for Proposition 4.5 and Proposition 4.6.

Other Valid Inequalities. Basic flow constraints (4.5) together with the advanced constraints on lifted edges (4.6)-(4.14) are sufficient for defining the set of feasible solutions of the lifted disjoint paths problem (4.3). Below, we present lifted flow inequalities specific to the lifted disjoint paths problem applied to MOT that help to improve the speed of our BLP solver. The inequalities depend on the fact that every node can be connected to maximally one node in each time frame. Therefore the number of lifted edges originating (or ending) in a given point and ending (resp. originating) in a specific time frame is at most one.

$$\begin{aligned} &\forall k, l \in [n_{\mathrm{R}}] \text{ s.t. } k > l,\ \forall v \in \mathcal{V}_l : \\ &\qquad \sum_{vu \in \breve{\mathcal{E}} : u \in \mathcal{V}_k} \breve{y}_{vu} \leq x_v \,, \end{aligned} \tag{4.18}$$

$$\begin{aligned} &\forall k, l \in [n_{\mathrm{R}}] \text{ s.t. } k < l,\ \forall w \in \mathcal{V}_l : \\ &\qquad \sum_{uw \in \breve{\mathcal{E}} : u \in \mathcal{V}_k} \breve{y}_{uw} \leq x_w \,. \end{aligned} \tag{4.19}$$

The number of constraints (4.18) and (4.19) is linear in the number of vertices. Therefore, we add them to our initial constraint set. This enables to reduce the search space for the branch-and-bound method in the early solver stages when only few constraints of type (4.9)-(4.14) have been added.

4.5 Separation

We solve the lifted disjoint paths problem (4.3) with the state-of-the-art integer linear program solver Gurobi [165]. Since there are exponentially many constraints of the form (4.9), (4.8), (4.12), (4.13), and (4.14), we do not add them initially. Instead, we start with constraints (4.5), (4.6), and (4.7) and find the optimal integer solution. If applied to MOT, we use in addition the constraints (4.18) and (4.19) to speed up the computation. In the separation procedures described below, we check if any of the advanced constraints are violated and add those that are to the active constraint set. We resolve the tightened problem and iterate until we have found a feasible solution to the overall problem (4.3).

Algorithms 4.1 and 4.2 describe the separation procedures for adding lifted path constraints (4.12), and lifted path-induced cut constraints (4.13) and (4.14). Since path constraints (4.9) and path-induced cut inequalities (4.8) are special cases of those above, they are also accounted for.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Separation for path inequalities. Algorithm 4.1 iterates over all active *st*-paths. For every path P^1, labels of all lifted edges connecting two vertices in $P^1_{\mathcal{V}}$ are inspected. If the lifted edge variable is zero, Algorithm 4.1 will extract a path in $\mathcal{G} \cup \breve{\mathcal{G}}$ connecting the endpoints and add the resulting lifted path inequality (4.12) to the active constraint set.

Algorithm 4.1: Separation for lifted path inequalities (4.12)

```
Define $\mathcal{E}^1 = \{e \in \mathcal{E} : y_e = 1\}$, $\mathcal{G}^1 = (\mathcal{V}, \mathcal{E}^1)$
for all $P^1 \in$ st-paths $(\mathcal{G}^1)$ do
    for all $\breve{y}_{vw} = 0 : v \in P^1_{\mathcal{V}} \wedge w \in P^1_{\mathcal{V}}$ do
        $P \leftarrow$ Extract_Path($P^1, v, w$)
        Add constr. (4.12) for $\breve{y}_{vw}$ with $P$.
    end for
end for
```

Separation for path-induced cut inequalities. Algorithm 4.2 iterates over all active *st*-paths. For every path P^1, lifted edges that start in $P^1_{\mathcal{V}}$ but do not end in $P^1_{\mathcal{V}}$ are inspected. If their label is one, Algorithm 4.2 will extract a subpath of P^1 for either (4.14) or (4.13) and add the respective inequality to the active constraint set.

Algorithm 4.2: Separation for lifted path-induced cut inequalities (4.13) and (4.14)

```
Define $\mathcal{E}^1 = \{e \in \mathcal{E} : y_e = 1\}$, $\mathcal{G}^1 = (\mathcal{V}, \mathcal{E}^1)$
for all $P^1 \in$ st-paths $(\mathcal{G}^1)$ do
    for all $\breve{y}_{vw} = 1 : v \in P^1_{\mathcal{V}} \wedge w \notin P^1_{\mathcal{V}}$ do
        if $\exists u \in P^1_{\mathcal{V}} : \breve{y}_{uw} = 0 \wedge vu \in \mathcal{R}$ then
            $P \leftarrow$ Extract_Path($P^1, v, u$)
            Add constr. (4.14) for $\breve{y}_{vw}$ with $P$.
        else
            $u \leftarrow$ last vertex of $P^1$ such that $uw \in \mathcal{R}$
            $P \leftarrow$ Extract_Path($P^1, v, u$)
            Add constr. (4.13) for $\breve{y}_{vw}$ with $P$.
        end if
    end for
end for
```

Complexity of separation. Both Algorithms 4.1 and 4.2 can be implemented efficiently such that they are linear[2] in $|\mathcal{E}^1|$ (*i.e.*, in the number of active edges of graph $\mathcal{G}$). In our implementation, we traverse all active *st*-paths from the end to the beginning and directly store correctly labeled lifted edges that originate on the already processed subpaths. These lifted edges can be used later as edges in $P_{\breve{\mathcal{E}}}$ in (4.12)-(4.14) or as $\breve{y}_{uw} = 0$ in (4.14).

---

[2] Assuming that the degree of each node can be upper bounded by a value δ independent of $|\mathcal{V}|$.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Algorithm 4.3: `Extract_path`(P^1, v, w)

```
$\dot{P} \leftarrow vw$-subpath of $P^1$
$P_{\mathcal{E}} \leftarrow \emptyset$
$P_{\breve{\mathcal{E}}} \leftarrow \emptyset$
for $j \in \dot{P}_{\mathcal{V}}$ from end of path to beginning do
    if $\exists$ edge $ij \in \breve{\mathcal{E}}$, $i \in \dot{P}_{\mathcal{V}}, \breve{y}_{ij} = 1$ then
        Add $ij$ to $P_{\breve{\mathcal{E}}}$
        Skip to node $i \in \dot{P}_{\mathcal{V}}$
    else
        Add $ij$ from $\dot{P}$ to $P_{\mathcal{E}}$
    end if
end for
output: $P = P_{\mathcal{E}} \cup P_{\breve{\mathcal{E}}}$
```

4.6 Complexity

Below, we show that the lifted disjoint paths problem (4.3) is $\mathcal{NP}$-hard. The following theorems state that even its restricted versions using only negative or only positive lifted edges are $\mathcal{NP}$-hard. The proofs use reductions from two known $\mathcal{NP}$-complete problems. Theorem 4.8 is proven by reduction from integer multicommodity flow [241], and Theorem 4.9 by reduction from 3-SAT [254].

We define $Y_{\mathcal{G}\breve{\mathcal{G}}}$ to be the set of all $(\mathbf{y}, \breve{\mathbf{y}}) \in \{0,1\}^{\mathcal{E}} \times \{0,1\}^{\breve{\mathcal{E}}}$ such that $(\mathbf{y}, \breve{\mathbf{y}})$ are feasible solutions of the lifted disjoint paths problem (4.3).

Integer multicommodity flow. The integer multicommodity flow problem is defined on a directed graph $\widetilde{\mathcal{G}} = (\widetilde{\mathcal{V}}, \widetilde{\mathcal{E}})$ with edge capacities $\mathfrak{c} \in \mathbb{N}^{\widetilde{\mathcal{E}}}$, and for all $i \in [k]$, source/sink pairs $s_i t_i$, edge flows $f^{(i)} \in \mathbb{N}^{\widetilde{\mathcal{E}}}$, and demands R_i.

The aim is to send k flows from their sources to their sinks such that the flows obey the edge capacities. Formally,

$$\sum_{i=1}^{k} f_e^{(i)} \leq \mathfrak{c}_e \qquad \forall e \in \widetilde{\mathcal{E}}\,, \tag{4.20}$$

$$\sum_{u:uv \in \widetilde{\mathcal{E}}} f_{uv}^{(i)} = \sum_{w:vw \in \widetilde{\mathcal{E}}} f_{vw}^{(i)} \qquad \forall i \in [k]\ \forall v \notin \{s_i, t_i\}\,, \tag{4.21}$$

$$\sum_{v:s_i v \in \widetilde{\mathcal{E}}} f_{s_i v}^{(i)} \geq R_i \qquad \forall i \in [k]\,. \tag{4.22}$$

Even *et al.* have shown [241] that the integer multicommodity flow problem is $\mathcal{NP}$-complete also in the case of unit capacity edges and two source sink pairs. Below we detail a construction that gives us a correspondence between edge-disjoint paths in $\widetilde{\mathcal{G}}$ and node-disjoint paths in a transformed graph $\mathcal{G}$. The lifted edges in the transformed graph will count how many units of flow go from sources to sinks.

Lemma 4.7. *There exists a polynomial transformation from any graph $\widetilde{\mathcal{G}}$ with source/sink pairs s_i, t_i, $i = 1, \ldots, k$ with demands R_i and unit capacity edges to a pair of graphs*

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

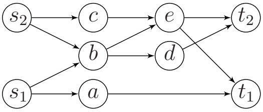


Figure 4.8: Integer multicommodity flow network transformation: Original graph.

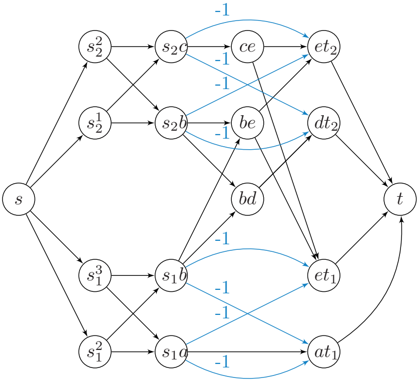


Figure 4.9: Integer multicommodity flow network transformation. Transformed graph from Figure 4.8 for flow demands $R_1 = 2, R_2 = 2$. Edges without label have cost 0.

$\mathcal{G}$ and $\breve{\mathcal{G}}$ with edge costs $\mathbf{q}$ and $\breve{\mathbf{q}}$, respectively such that there exists a feasible integer multicommodity flow in $\widetilde{\mathcal{G}}$ if and only if the lifted disjoint paths problem for $\mathcal{G}, \breve{\mathcal{G}}$ has objective

$$\min_{(\mathbf{y},\breve{\mathbf{y}})\in Y_{\mathcal{G},\breve{\mathcal{G}}}} \langle \mathbf{q}, \mathbf{y}\rangle + \langle \breve{\mathbf{q}}, \breve{\mathbf{y}}\rangle \leq -\sum_{i=1}^{k} R_i\,. \tag{4.23}$$

Proof. Without loss of generality, we consider these feasible flow sets $f_1, \dots, f_k$ where it holds $\forall i \in [k] : \sum_{s_i v \in \widetilde{\mathcal{E}}} f^{(i)}_{s_i v} = R_i$. Note that if the flow of commodity i is higher than its demand R_i, we can reduce it to R_i by removing the flow across one or more $s_i t_i$-paths in $\widetilde{\mathcal{G}}$ without violating other constraints.
We first detail the graph transformation (see Figures 4.8 and 4.9).

- For all edges $ij \in \widetilde{\mathcal{E}}$, add a vertex v_{ij} to $\mathcal{V}$.
- For each pair of vertices $v_{ij}, v_{jp} \in \mathcal{V}$, add an edge (v_{ij}, v_{jp}) to $\mathcal{E}$.
- Add vertices s and t to $\mathcal{V}$.
- Add to $\mathcal{V}$ vertices $s_i^1, s_i^2, \dots, s_i^{R_i}$ representing requirements of each commodity i.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

- For each vertex s_i^r, add an edge (s, s_i^r) to $\mathcal{E}$.
- For each pair of vertices s_i^r, v_{s_ij}, add edge (s_i^r, v_{s_ij}) to $\mathcal{E}$.
- For all $v_{pt_i} \in \mathcal{V}$ (representing an edge from p to t_i in $\widetilde{\mathcal{G}}$), add an edge (v_{pt_i}, t) to $\mathcal{E}$.
- For all pairs of vertices v_{s_ij} $v_{pt_i} \in \mathcal{V}$, add an edge (v_{s_ij}, v_{pt_i}) to $\check{\mathcal{E}}$. That is, the lifted edges connect all vertices representing edges from s_i in $\widetilde{\mathcal{G}}$ with vertices representing the edges to t_i in $\widetilde{\mathcal{G}}$.
- Cost function on base edges $\forall e \in \mathcal{E} : q_e = 0$.
- Cost function on lifted edges $\forall \check{e} \in \check{\mathcal{E}} : \check{q}_{\check{e}} = -1$.

An illustration of this construction can be seen in Figures 4.8 and 4.9. Note that the construction of $\widetilde{\mathcal{G}}$ in [241] allows $s_i = s_j$ for $i \neq j$. In this case, we still construct separate vertices for their incident edges in $\mathcal{G}$.

For arbitrarily chosen $r \in [R_i]$, every path $\widetilde{P} = (s_ik_1, k_1k_2, \dots, k_nt_i)$ in $\widetilde{\mathcal{G}}$ can be assigned to a path $P = (ss_i^r, s_i^r v_{s_ik_1}, v_{s_ik_1}v_{k_1k_2}, \dots, v_{k_nt_i}t)$ in $\mathcal{G}$ and vice versa. Note that such a path P saturates exactly one lifted edge $(v_{s_ik_1}, v_{k_nt_i})$. As we assume unit edge capacities, every feasible set of flow functions $f_1, \dots, f_k$ satisfying for all $i \in [k] : \sum_{s_iv \in \widetilde{\mathcal{E}}} f^i_{s_iv} = R_i$ defines a set of edge-disjoint paths from $s_1, \dots, s_k$ to $t_1, \dots, t_k$ in $\widetilde{\mathcal{G}}$. This set corresponds to a set of $\sum_{i=1}^k R_i$ st-paths in $\mathcal{G}$ whose edges and vertices are disjoint and where every path saturates exactly one lifted edge $v_{s_ij}v_{kt_i}$. Every lifted edge contributes with -1 to the total cost. So, this set of disjoint st-paths has total cost $-\sum_{i=1}^k R_i$.
Reversely, let us have a set of vertex- and edge-disjoint st-paths in $\mathcal{G}$ of size $\sum_{i=1}^k R_i$ where every path contains some $v_{s_ij}v_{kt_i}$-path as its subpath and therefore its cost is $-\sum_{i=1}^k R_i$. This set defines uniquely a set of feasible flow functions $f_1, \dots, f_k$.
So, there exist feasible functions $f_1, \dots, f_k$ satisfying $f_i = R_i$ for all $i \in [k]$ if and only if

$$\min_{(\mathbf{y},\check{\mathbf{y}}) \in Y_{\mathcal{G}\check{\mathcal{G}}}} \kappa(\mathbf{y}, \check{\mathbf{y}}) \leq -\sum_{i=1}^{k} R_i \,. \tag{4.24}$$

□

Theorem 4.8. *Lifted disjoint paths problem* (4.3) *with negative lifted edges only is $\mathcal{NP}$-hard.*

Proof. The $\mathcal{NP}$-complete integer multicommodity flow problem with unit edge capacities (finding a feasible solution) can be reduced in polynomial time to the lifted disjoint paths problem (4.3) with negative lifted edges only. The transformation is described in Lemma 4.7.

□

3-SAT. The boolean satisfiability problem (SAT) is a classical $\mathcal{NP}$-complete problem [254]. A transformation from its $\mathcal{NP}$-complete special version 3-SAT is commonly used for proving that a problem is $\mathcal{NP}$-hard or $\mathcal{NP}$-complete.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

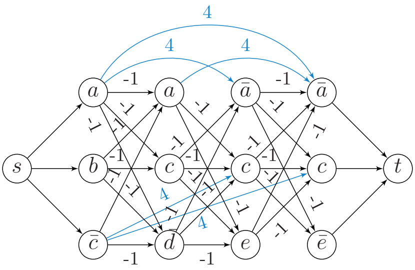


Figure 4.10: Reduction to lifted disjoint paths problem for 3-SAT formula $(a \vee b \vee \bar{c}) \wedge (a \vee c \vee \bar{d}) \wedge (\bar{a} \vee c \vee e) \wedge (\bar{a} \vee c \vee \bar{e})$.

Theorem 4.9. *Lifted disjoint paths problem* (4.3) *with positive lifted edges only is $\mathcal{NP}$-hard.*

Proof. Below, we detail a transformation from 3-SAT (see also Example 2.33) to the lifted disjoint paths problem with positive lifted edges only. For the transformation, it holds that a 3-SAT formula consisting of k clauses has a true assignment iff $\min_{(\mathbf{y},\breve{\mathbf{y}}) \in Y_{\mathcal{G}\breve{\mathcal{G}}}} \gamma(\mathbf{y},\breve{\mathbf{y}}) \leq -(k-1)$.

Let a 3-SAT problem containing k ordered clauses $C_1 \dots C_k$ be given. Each clause C_i consists of a conjunction of literals, which is either a variable a or its complement $\bar{a}$. We construct graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and $\breve{\mathcal{G}} = (\breve{\mathcal{V}}, \breve{\mathcal{E}})$ as follows.

- The graph $\mathcal{G}$ has k layers. Every layer corresponds to one clause. Each layer contains 3 vertices labeled with the literals in the corresponding clause. Specifically, for a variable a in clause C_i we associate node v_{ia}, analogously for a complemented variable $\bar{b}$ in clause C_i we associate node $v_{i\bar{b}}$.
- For every pair of vertices $v_{il_1} \in \mathcal{V}$ and $v_{(i+1)l_2} \in \mathcal{V}$ where $l_1 \neq \bar{l}_2$ add an edge $(v_{il_1}, v_{(i+1)l_2})$ to $\mathcal{E}$ and set $q_{(v_{il_1}, v_{(i+1)l_2})} = -1$.
- For every variable a and every pair of vertices $v_{ia}, v_{j\bar{a}} \in \mathcal{V}$ where $j > i+1$ add an edge $(v_{ia}, v_{j\bar{a}})$ to $\breve{\mathcal{E}}$ and set $\breve{q}_{(v_{ia}, v_{j\bar{a}})} = k$. Do so analogously for every pair of nodes $v_{i\bar{a}}$ and v_{ja}.
- Add an edge from s to all vertices corresponding to the first clause. And an edge to t from all vertices corresponding to the last clause.

An illustration of this construction can be found in Figure 4.10.

Every path $P \in st$-paths$(\mathcal{G})$ that has cost $-(k-1)$ saturates vertices labeled by non-contradicting literals. We can obtain a 3-SAT solution from P as follows. If $v_{ia} \in P_{\mathcal{V}}$, set variable $a := true$. If $v_{j\bar{b}} \in P$, set variable $b := false$. Variables not contained as labels of vertices in $P_{\mathcal{V}}$ can have arbitrary values.
Similarly, every solution of 3-SAT problem defines at least one path $P \in st$-paths$(\mathcal{G})$ that has cost $-(k-1)$.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

$\square$

4.7 Experiments

We conduct several experiments on MOT showing the merit of using lifted disjoint paths for the tracking problem. Below, we describe the problem construction, cost learning for base and lifted edges, pre-processing and post-processing steps, and report resulting performance.

4.7.1 Graph construction.

Two-step procedure. Due to the computational complexity of the problem, entire video sequences cannot be solved in a straightforward manner. In order to make the problem tractable, the following two-step procedure is used. In the first step, the solver is applied on graphs over person detections but only for small time intervals consisting of a few dozen video frames. The tracks resulting from the first step are used for extracting tracklets. In the second step, the solver is applied on newly created graphs $\mathcal{G}$ and $\breve{\mathcal{G}}$ where vertices correspond to the obtained tracklets. Edges and edge costs between tracklets are obtained by aggregating original edges resp. edge costs between person detections. The tracks resulting from the second step may be suboptimal with respect to the original objective function defined over person detections. Therefore, we identify points where splitting a track leads to an improvement of the original objective value and extract new tracklets from the divided tracks, similar to Section 3.3.2.4. Multiple iterations of the second step are performed until no improving split points are found in the output tracks. This two-step procedure improves the objective w.r.t. the original objective (4.3) in every iteration. Since there are only finitely many trackings, the procedure terminates finitely. In practice, only a few iterations are necessary.

Graph sparsification. For our experiments, we use edges between detections up to 2sec temporal distance (60 frames for sequences recorded at 30 fps). These long-range edges cause high computational complexity for the first step. In order to reduce it, we apply sparsification on both base and lifted graphs. For the base edges, we select for every $v \in \breve{\mathcal{V}}$ its K nearest (lowest-cost) neighbors from every subsequent time frame within an allowed time gap. Lifted edges with costs close to zero are not included, since they are not discriminative. Lifted edges connecting detections with high time gap are included more sparsely than lifted edges having lower time gaps. We use dense graphs in the second step.

Costs. Initially, in the first step, we set $c_v = 0$ for all vertices $v \in \mathcal{V}$. For the second step, where $\mathcal{V}$ represents tracklets, c_v is set to the cost of outputting tracklet v as a final trajectory. Specifically, c_v is the sum of costs of base edges between consecutive detections in the tracklet and the cost of lifted edges between all pairs of detections contained in the tracklet. The cost of a base edge between two tracklets is given by the cost of the original base edge connecting the last detection in the first tracklet

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

with the first detection in the subsequent tracklet. The cost of a lifted edge between two tracklets is obtained by summing up the costs of original lifted edges between detections contained in the tracklets. This ensures that the costs of the tracklet solution correspond to the costs of the original problem. We set the costs of all edges from the source node s and to the sink node t to zero. Setting of detection costs and in/out costs to zero reduces the number of hyperparameters that usually needs to be incorporated by other methods. Moreover, our method does not include temporal decay of edge costs since the formulation directly prefers short range base edges over the long-range ones. Cost definitions are provided in Section 4.7.3.

4.7.2 Pre-processing and post-processing

As is common for tracking-by-detection, we perform pre-processing and post-processing to compensate for detector inaccuracies.

Input filtering. Given a set of input detections derived from a detector, we follow the approach called Tracktor [58], a leading tracker for the MOT challenge, to reject false positive detections and to correct misaligned ones. For this, each input detection is sent through the regression and classification part of Tracktor's detector. In more detail, all tracking parts involved in the tracker Tracktor [58] are deactivated such that it only reshapes and eventually rejects input detections without assigning labels to them. Input detections are rejected if Tracktor's detector outputs a confidence score $\sigma_{\text{active}} \leq 0.5$.

Tracktor also applies a non-maxima-suppression on the reshaped input detections. We use the threshold $\lambda_{\text{new}} = 0.6$ (as defined by Bergmann *et al.* [58])

Inter- and extrapolation. Even if all input detections have been assigned to the correct identities by our solver, there might still be missing detections in case that a person has not been detected in some frames. We recover missing detections within the time range of a trajectory, which we denote as *interpolation*. Furthermore, we extend a trajectory in forward and backward direction, which we denote as *extrapolation*. To this end, we follow the approach of Tracktor [58] and apply their object detector to recover missing positions based on the visual information at the last known position. Finally, for sequences filmed by a static camera, we perform linear interpolation on the remaining gaps. Such sequences can be automatically detected using DeepMatching [177] on the regions outside detection boxes, see also Section 2.7.2.

To demonstrate the performance using traditional post-processing, we also evaluate our tracker using only linear interpolation as post-processing in all sequences in Section 4.7.7.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

4.7.3 Cost learning

Costs for base edges $\mathcal{E}$ and lifted edges $\breve{\mathcal{E}}$ are computed equally[3], since they both indicate whether two detections are from the same object or not. For an edge $e = vw$, we denote with $d_{\text{wi}}(v)$ the detection width corresponding to node v.

Visual cues. Two different appearance features are exploited: Given two detections, the *re-identification* descriptor utilizes global appearance statistics, while the *deep-matching* descriptor relies on fine-grained pixel-wise correspondences.

To exploit global appearance statistics, the re-identification network [178] introduced in Section 2.7.2.2 is employed, which we train on the MOT17 training set [44] and additional re-identification datasets [181, 186, 255]. We extract from the network for each edge $e = vw$ a similarity value $\psi_{\text{re-id}}(e) \in [0, 1]$ such that the higher the value, the more likely v and w correspond to the same person.

Given a valid match, *i.e.*, an edge $e = vw$ connecting detections of the same person and an invalid match, *i.e.*, an edge $e' = vw'$ connecting v with a detection w' of another person, the network usually creates similarity scores such that $\psi_{\text{re-id}}(e) > \psi_{\text{re-id}}(e')$. However, we observed that the range of similarity scores $\{\psi_{\text{re-id}}(e) \mid e \text{ valid match}\}$ on the MOT17 dataset is rather large. To increase precision and better reflect uncertainty of a connection, a novel feature transformation is proposed: Each value is normalized by the similarity scores appearing within the involved time frames. To this end, we define for an edge $e = vw$ with $v \in \mathcal{V}_{\mathfrak{f}_1}$ and $w \in \mathcal{V}_{\mathfrak{f}_2}$ the normalization values $\psi^{\rightarrow}_{\text{norm}}$ and $\psi^{\leftarrow}_{\text{norm}}$ in temporal forward and backward direction, respectively:

$$\psi^{\rightarrow}_{\text{norm}}(e) \coloneqq \frac{\psi_{\text{re-id}}(vw)}{\max_{v' \in \mathcal{V}_{\mathfrak{f}_1}} \psi_{\text{re-id}}(v'w)} \in [0,1], \quad \psi^{\leftarrow}_{\text{norm}}(e) \coloneqq \frac{\psi_{\text{re-id}}(vw)}{\max_{w' \in \mathcal{V}_{\mathfrak{f}_2}} \psi_{\text{re-id}}(vw')} \in [0,1]\,. \tag{4.25}$$

The features $\psi^{\rightarrow}_{\text{norm}}(e)$ and $\psi^{\leftarrow}_{\text{norm}}(e)$ thus indicate if an edge e is the best selection for the involved detections among the respective frames. Since the normalization exploits information from entire frames instead of just two detections, we call them *global context normalization* features. Finally, for each edge e, the following features are used:

$$(\psi^{\rightarrow}_{\text{norm}}(e), \psi^{\leftarrow}_{\text{norm}}(e), \psi^{\rightarrow}_{\text{norm}}(e)\psi_{\text{re-id}}(e), \psi^{\leftarrow}_{\text{norm}}(e)\psi_{\text{re-id}}(e)) \in [0,1]^4\,. \tag{4.26}$$

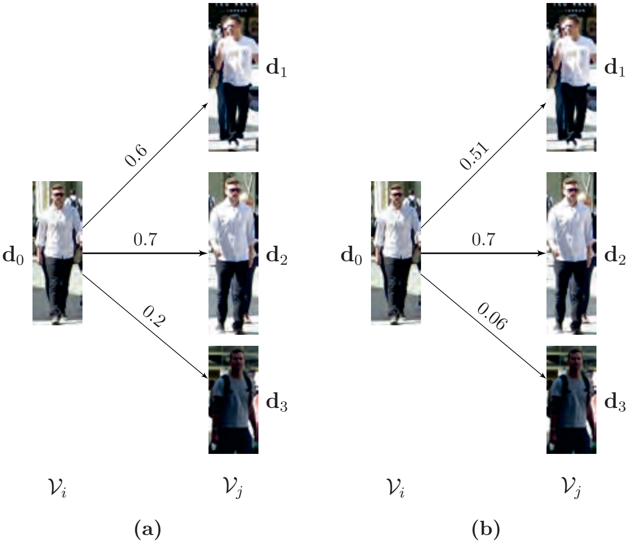
The global context normalization features are visualized in Figure 4.11.

Our second visual cue utilizes DeepMatching (DM) [256], introduced in Section 2.7.2.1, to establishes pixelwise correspondences between two images. We apply DM between boxes in two images and compute the DM intersection over union [55, 63] (see Section 3.4.1) w.r.t. the whole detection boxes and on five sub boxes (left/right, upper/middle/lower part), which creates six features per box pair as illustrated in Figure 4.12. In addition, we measure for all DM points in a given sub box whether their matched endpoints are in the corresponding sub box again or not, represented by an indicator value. By using the left/right and the upper/middle/lower part subdivision, this gives two additional error measures for deviation in x and y-directions. Thus, in total, we obtain a feature

---

[3] By using the same costs for base and lifted edges, more weight is given to the base graph, compare also Figure 1.11.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.



(a) **(b)**

Figure 4.11: Illustration of global context normalization on the MOT17 dataset. (a) Depicted are the similarities $\psi_{\text{re-id}}(e)$ between all detections of two frames. The similarity score of the valid match $(\mathbf{d}_0, \mathbf{d}_2)$ is highest among all similarities from $\mathbf{d}_0$ to a detection of $\mathcal{V}_j$. Due to a change in the pose and background, the value is slightly low. Most importantly, the wrong match $(\mathbf{d}_0, \mathbf{d}_1)$ yields a similarly high value. (b) The global context normalization $\psi^{\rightarrow}_{\text{norm}}(e)\psi_{\text{re-id}}(e)$ is much more discriminative. Some parts of the image are taken from the MOT16 dataset [44].

vector $\boldsymbol{\psi}_{\text{DM}}(e) \in [0,1]^8$. In order to assess the reliability of DM features, the coverage of each box and its sub boxes by matching points is computed. Now for each pairing (detection boxes or sub boxes), the smaller of the two coverage values is chosen, which creates feature $\boldsymbol{\rho} \in [0,1]^6$.

Motion constraints. We penalize for improbable motions by comparing the maximal displacement of DM endpoints within a considered sequence with the displacements of detection boxes. Assignment hypotheses of pairs of boxes representing improbable motions are penalized with a large cost.

Spatio-temporal cues. Our spatio-temporal cues utilize a simple motion compensation by computing the median DM displacement between correspondences of the background.

We assume a linear motion model, similar to Ristani *et al.* [78] and penalize deviations of detections from the estimated motion trajectory. This enforces spatio-temporal consistency of detections within one trajectory. Furthermore, we penalize improbable large person movements by relating velocities (in pixels per seconds) in horizontal

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

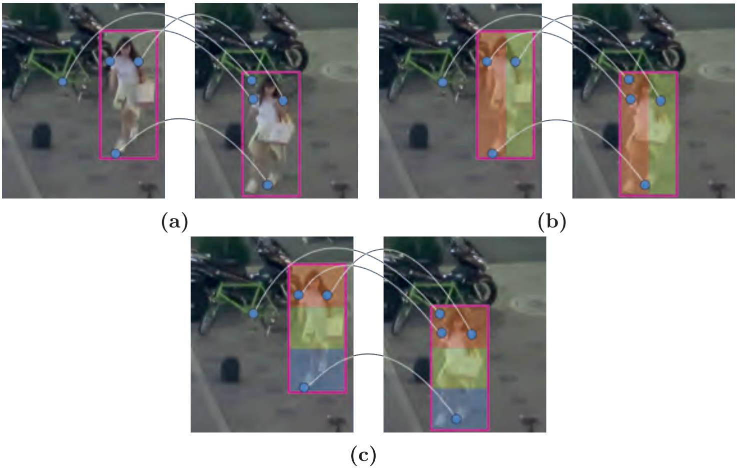


Figure 4.12: Pairwise features derived from DeepMatching. (a) DM intersection over union is computed between detection boxes. In addition, detection boxes are subdivided vertically (b) and horizontally (c) so that DM intersection over union is also computed between equally colored sub boxes. This considers not only the presence of DM endpoints in corresponding detection boxes but also that they are spatially consistent. Some parts of the image are taken from the MOT16 dataset [44].

direction (ν_x) to box width wi_v and wi_u of nodes v and u, respectively: $\psi_{\text{trans}}(e) = \log(\nu_x(e)/\min\{\mathrm{wi}_v, \mathrm{wi}_u\})$.

Fusion of input features. We construct a neural network consisting of fully-connected layers, batch normalization, and relu units, taking the above described features and time differences as input. The final layer uses a sigmoid activation function for producing a score in $[0,1]$. The output of the network represents the likelihood for an assignment hypothesis of two detection boxes. It is converted to costs using the log-odd-function, see Eq. (2.11).

In more detail, considering one assignment hypothesis represented by an edge $e = vw$, the DeepMatching densities $\boldsymbol{\rho} \in [0,1]^6$ as well as the temporal distance t between the corresponding detections v and w serve as a confidence score for the remaining input features. They describe which of the input features are reliable metrics for a given assignment hypothesis, but they do not provide any information that can be used to decide the validity of a considered assignment hypothesis. We transform the density features non-linearly and denote them together with the temporal distance (of maximal 2 seconds) as control features $\mathcal{C}(e) := (\log(\boldsymbol{\rho}), t) \in \mathbb{R}^6 \times [0,2]$. The remaining described features of this section are denoted as $\boldsymbol{\psi}(e) \in [0,1]^n$.

One plausible architecture is to use a convex combination of the input features such that the coefficients depend on the control features. To this end, let $\alpha_i(\mathcal{C}(e), \mathbf{W}_{\alpha_i})$ for

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

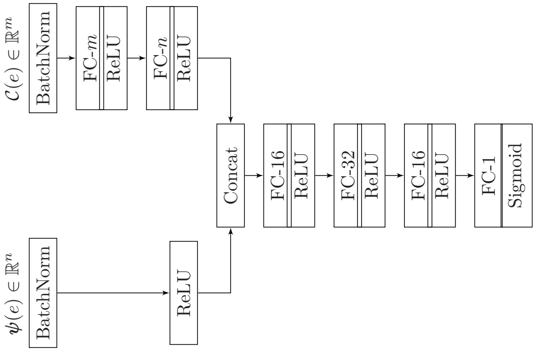


Figure 4.13: The architecture of the employed edge classifier. FC-i denotes a fully-connected layer with i output nodes. Using a concatenation with subsequent fully connected layers, m control features and n input features are fused.

$i \in [n]$ denote a neural network with the control features as input and $\mathbf{W}_{\alpha_i}$ as learnable weights. Furthermore, let $\beta_i(\psi(e)_i, \mathbf{W}_{\beta_i})$ for $i \in [n]$ be a neural network applied to i-th feature of $\boldsymbol{\psi}(e)$ with learnable weights $\mathbf{W}_{\beta_i}$.

The input features and control features can then be fused via

$$\sum_{i=1}^{n} \alpha_i(\mathcal{C}(e), \mathbf{W}_{\alpha_i}) \beta_i(\psi(e)_i, \mathbf{W}_{\beta_i}) \tag{4.27}$$

such that

$$\sum_{i=1}^{n} \alpha_i(\mathcal{C}(e), \mathbf{W}_{\alpha_i}) = 1\,. \tag{4.28}$$

To ensure stable training, (4.27) should be applied to a sigmoid function and trained using binary cross-entropy loss.

Nonetheless, our tracker implementation employs a neural network based mainly on a combination of relu units and fully connected layers, which performed slightly better, still sharing the idea of separating the input into control features and input features. The detailed architecture is depicted in Figure 4.13.

Training details. Training of the neural network is performed directly on the (preprocessed) input detections. Labels are retrieved by assigning each detection to the best fitting ground truth bounding box. Detections with ambiguous assignments are ignored within the training phase.

In order to train the edge classifier, special care has to be taken as the training set is highly imbalanced. The number true negative edges (pairs of detections that do not belong to the same person) clearly dominates the number of true positive edges (pairs of detections belonging to the same person). To address this issue, the network is trained on a randomly sampled subset of all possible edges such that the ratio of true positive edges and true negative edges per time distance between the end nodes of the edges

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

remains fixed. The maximal temporal distance of an edge is set to 2 seconds, allowing to recover persons even after long occlusions.

The weights of the fusion network are optimized according to the binary cross-entropy loss. We employ stochastic gradient descent with the learning rate set to 10^{-2} and Nesterov momentum set to 0.9, for a total of 10 epochs. Training and inference are performed using Pytorch 1.3 on an Nvidia RTX 2080 Ti.

4.7.4 Implementation details on the lifted disjoint paths solver

The solver for the lifted disjoint paths problem is implemented in C++ and builds upon Gurobi 7.5. All subsequent experiments were conducted on a machine with a 6-Core Intel 2.00GHz CPU and 128 GB RAM.

4.7.5 Experiment setup

In order to assess the suitability of the proposed lifted disjoint paths formulation for MOT, extensive experiments are conducted on the datasets MOT15/16/17 [43, 44]. We perform analysis and parameter tuning for our tracker on the MOT17 train set, even when the tracker is applied to the MOT15 sequences, to ensure that the tracker is not prone to overfitting. We follow the MOT challenge protocol and use the detections provided by the respective benchmarks. All experiments on the training set are evaluated using a leave-one-out cross-validation. This includes all of the training procedures, in particular also the training of the re-identification network. The tracking quality produced by the proposed method is measured in terms of the event-based and identity-based metrics of Section 2.7.4. In the following, we denote by *Lifted Disjoint Paths Tracker* (Lif_T) the proposed tracker of this chapter.

Table 4.1: Assignment quality of our solver without interpolation or extrapolation on the MOT17 train set with different maximal time gaps in seconds. Rows 1, 3, 5, and 7 show the results by our solver, rows 2, 4, 6, and 8 show the maximally achievable bounds with admissible assignment hypotheses up to the specified time gap. Bold numbers represent the best values per row.

	0.3s	0.5s	1s	1.5s	2s	∞
MOTA[%] (ours) ↑	52.6	52.7	52.8	52.8	**52.8**	-
MOTA[%] (optimal) ↑	53.0	53.1	53.3	53.3	**53.4**	**53.4**
IDF1[%] (ours) ↑	55.7	57.8	61.8	63.8	**64.3**	-
IDF1[%] (optimal) ↑	56.0	58.6	63.2	65.7	66.8	**69.9**
IDP[%] (ours) ↑	79.8	82.9	88.5	91.4	**92.1**	-
IDP[%] (optimal) ↑	80.4	84.2	90.8	94.3	95.9	**100.0**
IDR[%] (ours) ↑	42.7	44.5	47.4	49.0	**49.4**	-
IDR[%] (optimal) ↑	42.9	45.0	48.5	50.4	51.3	**53.4**

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

4.7.6 Benefit of long-range edges

We investigate the importance of using long-range information for MOT. To this end, we apply our proposed tracker Lif_T on the MOT17 training sequence with varying maximal time gap, for which base and lifted edges are created between nodes. In order to assess the influence of the time gap on the tracking quality, we measure the *assignment* quality in terms of the MOTA and IDF1 metrics without performing any interpolation or extrapolation. To assess how well the *assignment part* is solved by our tracker, we compute the maximum achievable metrics given the filtered input detections and admissible assignment hypotheses within maximal time gaps. A detailed description of how we obtain the optimal assignments is given below. From the result in Table 4.1 we see essentially constant MOTA scores. This is due to the fact that selecting correct connections does not change MOTA significantly except after interpolation and extrapolation (which we have excluded in Table 4.1). However, we see a significant improvement in the IDF1 score, which directly penalizes wrong connections. Here, long-range edges help greatly. Moreover, both metrics, ID precision and ID recall, clearly increase with increasing time gap. This shows that improvements by incorporating more temporal information come from using longer skip edges (impact on IDR) but most importantly, precision increases greatly. This means that ID switches are avoided thanks to lifted edges. Furthermore, the experiment shows that our designed features together with the lifted disjoint paths formulation (4.3) are well-suited for the MOT problem delivering nearly optimal assignments.

Optimal assignment of detections. We elaborate on the details to obtain the optimal assignments. We start with the pre-processed input detections, according to Section 4.7.2. For each frame, we compute the intersection over union between the detections and ground truth boxes of the respective frame, which forms a weighted bipartite graph. Edges with a corresponding intersection over union below 0.5 are removed. Then, we use Hungarian matching to find a maximum-weighted matching. Unmatched detections are considered as false positives, while matched detections are assigned the corresponding ground truth label. Thus, we obtain the trajectories on the input detections using the optimal assignment. Finally, depending on the time threshold of Table 4.1, trajectories are synthetically splitted at skip-edges longer than the specified threshold such that each segment obtains a unique ID.

4.7.7 Ablation study on post-processing methods.

Solving the proposed lifted disjoint paths problem establishes the assignment of input detections to object identities very close to the best possible assignment (Section 4.7.6).

To localize tracked objects also in those frames in which the object detector failed to detect them, some trackers apply an additional object detector on these frames based on the available input detections. This can be seen as performing interpolation and extrapolation if viewed from the perspective of data association in a tracking-by-detection framework, *e.g.*, see Bergmann *et al.* [58]. As a result, improvements can be achieved from extending trajectories to image areas without input detections by applying a very accurate object detector.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 4.2: Ablation study on interpolation and extrapolation, evaluated on the MOT17 train set. SI = spatial interpolation only on sequences filmed by a static camera, SI* = spatial interpolation on all sequences, VI = visual interpolation, VE = visual extrapolation. *Assignment* and *Assignment (optimal)* denote the results of the lifted disjoint paths problem and the optimal assignment, as reported in Section 4.7.6 given 2sec time gap. Note that Tracktor's object detector is fine-tuned on MOT17Det. In our experiments, this resulted in bigger improvements on the MOT17 training set than on the test set, compare Table 4.3.

Method	MOTA[%] ↑	IDF1[%] ↑
Assignment	52.8	64.3
Assignment (optimal)	53.4	66.8
Assignment+SI	57.8	67.6
Assignment+SI*	59.5	68.9
Assignment+VI	59.6	68.5
Assignment+VI+VE	65.7	71.5
Assignment+VI+VE+SI	**67.0**	72.4

To make our tracking performance comparable to the results of other trackers, we follow this strategy and employ interpolation and extrapolation, based on the method Tracktor by Bergmann *et al.* [58]. During the interpolation and extrapolation, detections coming from the lifted disjoint paths solver are preserved. In particular, the detections are not rejected, reshaped, neither are their labels changed by Tracktor. Instead, we apply Tracktor to recover further locations of an object in those frames in which detections of the object were missing. The procedure is based on the trajectories obtained from the lifted disjoint paths solver. Note that our adaption ignores additional, unassigned input detections, whereas the original implementation [58] of Tracktor fuses the detections coming from Tracktor's detector with detections provided by the dataset.

Table 4.2 reports the influence of employing interpolation and extrapolation. The first two rows repeat values from Table 4.1 given the maximal 2s time gap. Since our solver produces nearly optimal data assignment with respect to the used input detections, further improvements can only be achieved by applying interpolation and extrapolation on the tracks obtained from the solver.

We compare visual interpolation (VI), visual extrapolation (VE), both using the method of [58], and spatial interpolation (SI). For SI, we employ linear interpolation based solely on the geometric bounding box information. The interpolation SI is applied only to sequences with a fixed camera in order to guarantee robust approximations. Still, the improvements by Assignment+SI over the baseline is evident. Especially the MOTA metric, which measures mainly the coverage of objects by detections, improves by about 10%. We also evaluate spatial interpolation for all sequences (SI*), which improves the tracker further to 59.5% MOTA and 68.9% IDF1. However, performing spatial interpolation on sequences with moving cameras can lead to error propagation.

Conversely, the visual interpolation based on Bergmann *et al.* [58] can be applied robustly to all sequences, but only in those situations in which an object is visible. Accordingly, the method Assignment+VI further improves over the baseline.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Recovering the position of tracked objects also outside of the time range of its computed trajectory (Assignment+VI+VE) further helps to improve the tracking accuracy, enhancing MOTA by about 20% and IDF1 by about 10% IDF1, as VE extends computed trajectories, thereby achieving longer identity consistencies.

Finally, we employ spatial interpolation on the remaining cases where detections are missing and the objects are fully occluded (Assignment+VI+VE+SI), resulting in a slight improvement over Assignment+VI+VE. The best-performing method according to Table 4.2, Assignment+VI+VE+SI, is thus used to evaluate the proposed tracker on the MOT15, MOT16, and MOT17 test set (see Section 4.7.10).

According to Table 4.2, the impact of the post-processing using Tracktor seems to be very high. We conjecture this might be due to the fact that Tracktor's object detector is trained on the MOT17Det train set (which are the detections of the MOT17 train set), leading to some degree of overfitting on the MOT17 training set. Note that Tracktor is not trained on the MOT17 tracking ground truth, so that it is still regarded as a meaningful validation procedure [58]. Our conjecture is supported by the conducted experiments presented in Section 4.7.10 which show that the post-processing method applied to the test set only slightly improves the results of the proposed lifted disjoint paths approach.

4.7.8 Accuracy of the fusion network

The performance of a tracking method depends highly on the accuracy of the edge classifier. Therefore, we report the evaluation of the edge classifier on all training sequences of the filtered MOT17 train set used for Table 4.4. Together with Table 4.5 and Table 4.3, it shows that improvements in tracking features directly correlate to high-quality tracking results thanks to the proposed solver. While Table 4.4 shows very good performance of the edge classifier, a powerful graph model and solver is still crucial to obtain high-quality tracking results. Even small errors (we observed 5% maximal error) in the edge classifier can cause many errors in the tracking results if an unsuitable procedure is used. Also, note that for training the edge classifier, detections with ambiguous assignments to the ground truth boxes were ignored. Thus, these potentially difficult cases are excluded in the evaluation of the edge classifier. Especially the interpolation and extrapolation are prone to error propagation once a single identity switch has been created, which heavily affects, among others, the IDF1 score. Our lifted disjoint paths formulation can be advantageous since lifted edges aggregate multiple edge classifiers, which can correct individual wrong classifications of single edges.

4.7.9 Qualitative evaluations

We provide a qualitative comparison of Lif_T with the best performing competing method Tracktor [58] in Figure 4.14. It shows that Lif_T successfully incorporates long-range information to produce long-term consistent trajectories. In contrast, object occlusions cause Tracktor to create new trajectories. Additional visual results depicted in Figure 4.15 show that Lif_T produces long-term consistent trajectories.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

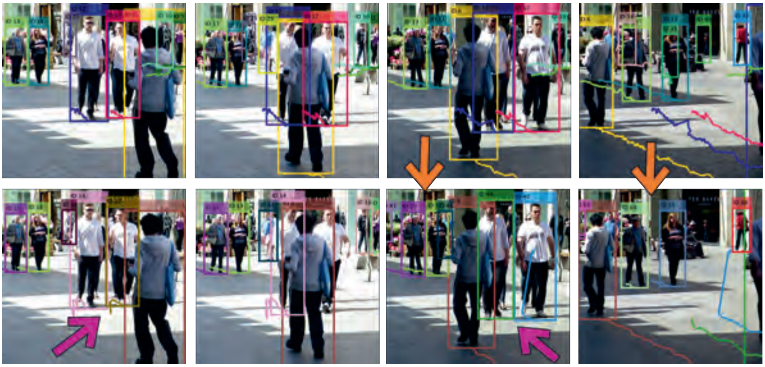


Figure 4.14: Comparison of Lif_T (top row) with Tracktor [58] (bottom row) on MOT17-08-SDP. Tracktor is not able to exploit long-term temporal interactions leading to ID switches and missing detections, as highlighted in the image. In contrast, Lif_T produces correct trajectories despite heavy occlusions.

4.7.10 Benchmark evaluations

We compare the tracking performance of Lif_T on the MOT15, MOT16, and MOT17 benchmarks with all trackers listed on the MOTChallenge which have been peer-reviewed and correspond to published work. Table 4.3 shows the tracking performance of the proposed tracker together with the best 5 performing trackers at the time this work was published [50], accumulated over all sequences of the respective benchmarks. The evaluations show that the proposed method outperforms all competing tracking methods by a large margin on all considered benchmarks. On MOT17, the MOTA score is improved from 53.5% to 60.5% and the IDF1 score from 52.3% to 65.6%, which corresponds to an improvement of 13% in terms of MOTA and almost 25% in terms of the IDF1 score, indicating the effectiveness of the lifted edges. We observe similar improvements across all three benchmarks. These results reflect the near-optimal assignment performance observed on the MOT17 train set in Section 4.7.6. On average, the BLP solver needs 26.6 min. per sequence. Detailed runtimes are available in Table 4.5. In addition, the evaluations on the MOT15, MOT16, MOT17 test sets as well as the MOT17 train set per sequence are provided in Table 4.5.

Finally, to analyze the impact of the post-processing on the test set, an additional tracker, *Lifted Disjoint Paths Tracker using simple linear interpolation* (Lif_TsimInt), is evaluated on all test sequences of the MOT15, MOT16, and MOT17 benchmarks. This tracker uses linear interpolation between detections of a trajectory for all sequences, which corresponds to Assignment+SI* in Table 4.2. Comparing the performance gain on the test set using Lif_T instead of Lif_TsimInt with the performance gain on the training set (Table 4.2), we see that indeed, the impact of the post-processing on the test set is significantly lower. We conclude that while the post-processing improves the tracking performance, the main performance of Lif_T is due to the proposed contributions. Lif_T is able to achieve near-optimal results with respect to the input

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

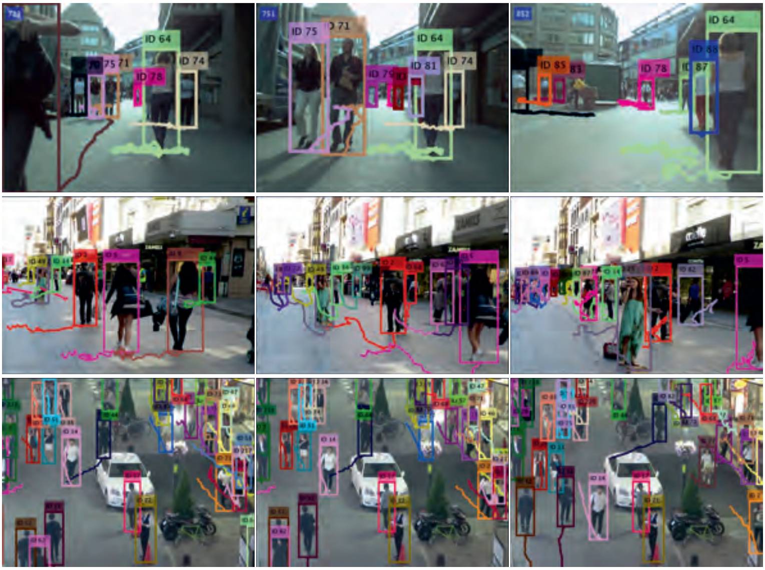


Figure 4.15: Tracking results of Lif_T on MOT17-06-SDP (first row), MOT17-08-SDP (second row), and MOT-03-SDP (third row).

detections. Applying interpolation and extrapolation according to the post-processing method further improves the results and makes it conceptually comparable to Tracktor. With post-processing on our computed data association, we improve over Tracktor by 25%. We argue that solving the data association accurately is important to obtain a final high-quality result after post-processing.

4.8 Conclusion

The method presented in this chapter demonstrates that a powerful data association model with a global optimal solver helps to compensate for erroneous inputs (Section 1.3). The novel LDP formulation extends the well-established disjoint paths problem, which is a natural model for MOT as trajectories are represented by vertex-disjoint paths. By the integration of lifted edges, path connectivity priors can be modeled so that the model becomes more expressive. The evaluations have shown that incorporating long-range connections using the LDP formulation improves tracking results considerably. Using edges connecting nodes over maximal 2 seconds instead of 0.3 seconds improves the precision by more than 15% in terms of IDP. This is in accordance with the concept of lifted edges that help to avoid ID switches. In the same setup, the recall is also improved by more than 15% in terms of IDR. Here, lifted edges help to re-identify persons that have been occluded. Consequently, the IDF1 score is improved substantially by more than 15% in the conducted experiment on the MOT17 train set.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 4.3: We compare our tracker Lif_T with the five best performing competing solvers w.r.t. MOTA from the MOT challenge. Tracktor [58], JBNOT [101], FAMNet [88], eTC [257], eHAF [258], NOTA [259], HCC [260], KCF [261], AP_HWDPL_p [262], STRN [82] and AMIR15 [91]. In addition, we compare the results to our tracker Lif_TsimInt that uses only a simple interpolation method (linear interpolation) as post-processing in all sequences. We outperform competing solvers on most metrics on all three MOT Challenge benchmarks, using Lif_T and Lif_TsimInt. The result table was retrieved when the work [50] was submitted (07.01.2020). Arrows indicate whether low or high metric values are better.

	Method	MOTA[%] ↑	IDF1[%] ↑	MT[%] ↑	ML[%] ↓	FP ↓	FN ↓	IDS ↓
MOT17	Lif_T (ours)	**60.5**	**65.6**	27.0	33.6	14966	**206619**	1189
	Lif_TsimInt (ours)	58.2	65.2	**28.6**	33.6	16850	217944	**1022**
	Tracktor17	53.5	52.3	19.5	36.6	**12201**	248047	2072
	JBNOT	52.6	50.8	19.7	35.8	31572	232659	3050
	FAMNet	52.0	48.7	19.1	**33.4**	14138	253616	3072
	eTC17	51.9	58.1	23.1	35.5	36164	232783	2288
	eHAF17	51.8	54.7	23.4	37.9	33212	236772	1834
MOT16	Lif_T (ours)	**61.3**	**64.7**	**27.0**	34.0	4844	**65401**	389
	Lif_TsimInt (ours)	57.5	64.1	25.4	**34.7**	4249	72868	**335**
	Tracktor16	54.4	52.5	19.0	36.9	**3280**	79149	682
	NOTA	49.8	55.3	17.9	37.7	7248	83614	614
	HCC	49.3	50.7	17.8	39.9	5333	86795	391
	eTC	49.2	56.1	17.3	40.3	8400	83702	606
	KCF16	48.8	47.2	15.8	38.1	5875	86567	906
2D MOT15	Lif_T (ours)	**52.5**	**60.0**	**33.8**	**25.8**	6837	**21610**	730
	Lif_TsimInt (ours)	47.2	57.6	27.0	29.8	7635	24277	**554**
	Tracktor15	44.1	46.7	18.0	26.2	6477	26577	1318
	KCF	38.9	44.5	16.6	31.5	7321	29501	720
	AP_HWDPL_p	38.5	47.1	8.7	37.4	**4005**	33203	586
	STRN	38.1	46.6	11.5	33.4	5451	31571	1033
	AMIR15	37.6	46.0	15.8	26.8	7933	29397	1026

The proposed tracker of this chapter outperforms the state of the art on all considered datasets, *i.e.*, MOT15, MOT16, and MOT17, by a large margin when the work was published [50]. For instance on MOT17, the MOTA score is improved by about 12% and the IDF1 score by about 25%.

While the proposed method is able to compute trajectories for all sequences of MOT15, MOT16, and MOT17, the necessary runtime as provided in Table 4.5 shows that the instance size is a limiting factor. Long sequences containing many detections (*e.g.*, MOT16-03 and MOT16-04) already require considerable runtime.

Overall, the novel LDP formulation is an expressive model for MOT. Presented features are accurate and robust also for long-term temporal connections. The proposed solver is, to our knowledge, the first to solve an implicit HO-MOT model using global optimization in practice. It allows to accurately exploit the information provided by the input detections, in particular long-range interactions without introducing errors by the

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 4.4: Performance metrics on the edge classifier. The performance is measured in terms of the *Accuracy* (Acc), *Precision* (Prec), *True Positive Rate* (TPR), and *True Negative Rate* (TNR). The arrows indicate that higher metric values are better.

Sequence	Acc[%] ↑	Prec[%] ↑	TPR[%] ↑	TNR[%] ↑
MOT17-02-DPM	100	99	100	100
MOT17-04-DPM	100	98	99	100
MOT17-05-DPM	95	95	100	99
MOT17-09-DPM	100	98	98	100
MOT17-10-DPM	100	99	99	100
MOT17-11-DPM	100	100	99	100
MOT17-13-DPM	99	97	96	100
MOT17-02-SDP	100	96	100	100
MOT17-04-SDP	100	98	98	100
MOT17-05-SDP	99	92	100	98
MOT17-09-SDP	97	81	99	97
MOT17-10-SDP	99	94	97	100
MOT17-11-SDP	100	99	99	100
MOT17-13-SDP	99	90	96	99
MOT17-02-FRCNN	100	98	100	100
MOT17-04-FRCNN	100	97	99	100
MOT17-05-FRCNN	99	94	100	100
MOT17-09-FRCNN	99	97	98	100
MOT17-10-FRCNN	99	95	98	100
MOT17-11-FRCNN	100	99	99	100
MOT17-13-FRCNN	99	90	95	99

solver. As a consequence, the method significantly improves over competing methods. Furthermore, the experiments show that for the MOT challenge datasets, nearly optimal data association performances are reached. We conjecture that further improvements would have to come from better detectors, better interpolation and extrapolation, and more powerful solvers for our formulation to take into account even longer time-gaps. Our polyhedral work offers the basis for writing such more powerful solvers.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Table 4.5: Per sequence evaluation of our tracker Lif_T. We also provide the solver time spent to solve the corresponding lifted disjoint paths problem instance (STime) in seconds. Arrows indicate whether low or high metric values are better. Tracking results on the test sets were evaluated by the MOTChallenge server `https://www.motchallenge.net` .

	Sequence	MOTA[%]↑	IDF1[%]↑	MT[%]↑	ML[%]↓	FP↓	FN↓	IDS↓	STime↓
MOT17-Train	MOT17-02-DPM	40.5	50.3	21	47	19	11017	26	127
	MOT17-04-DPM	69.9	73.9	49	27	298	13986	38	1521
	MOT17-05-DPM	58.2	67.0	23	30	40	2824	27	36
	MOT17-09-DPM	72.9	71.6	54	4	58	1370	15	59
	MOT17-10-DPM	67.4	70.2	46	14	106	4043	39	173
	MOT17-11-DPM	67.3	73.9	32	35	55	3017	11	115
	MOT17-13-DPM	63.6	67.2	41	33	64	4127	43	59
	MOT17-02-FRCNN	47.4	57.2	24	36	89	9656	26	229
	MOT17-04-FRCNN	67.5	74.1	46	25	98	15310	29	1535
	MOT17-05-FRCNN	60.2	68.9	26	27	73	2651	30	92
	MOT17-09-FRCNN	71.5	72.9	54	4	54	1451	10	51
	MOT17-10-FRCNN	73.2	76.2	58	4	270	3096	73	398
	MOT17-11-FRCNN	73.1	78.8	43	24	82	2436	18	133
	MOT17-13-FRCNN	77.1	75.8	62	10	203	2394	73	388
	MOT17-02-SDP	55.0	61.3	26	26	65	8236	52	586
	MOT17-04-SDP	77.7	81.8	56	16	243	10296	49	4133
	MOT17-05-SDP	64.0	69.5	31	17	105	2351	33	80
	MOT17-09-SDP	73.0	73.0	54	4	69	1356	12	127
	MOT17-10-SDP	75.0	78.6	61	4	349	2759	105	756
	MOT17-11-SDP	74.4	78.4	48	19	115	2277	27	198
	MOT17-13-SDP	70.8	71.4	56	21	200	3150	55	364
	MOT17-Train	67.0	72.4	41	22	2655	107803	791	11430
MOT17-Test	MOT17-01-DPM	48.3	58.1	33	46	68	3258	10	38
	MOT17-03-DPM	73.3	70.1	55	11	3560	24276	160	24311
	MOT17-06-DPM	58.1	64.7	27	35	178	4728	28	113
	MOT17-07-DPM	44.4	52.3	12	35	155	9176	60	297
	MOT17-08-DPM	34.7	47.4	24	49	254	13507	32	146
	MOT17-12-DPM	48.3	62.3	20	45	35	4437	11	68
	MOT17-14-DPM	36.1	48.8	7	47	268	11449	91	323
	MOT17-01-FRCNN	47.7	58.1	33	42	246	3119	7	79
	MOT17-03-FRCNN	72.2	71.8	48	11	2664	26277	124	11678
	MOT17-06-FRCNN	60.4	63.7	31	27	279	4358	32	203
	MOT17-07-FRCNN	44.0	54.9	13	33	279	9110	63	281
	MOT17-08-FRCNN	31.9	43.3	23	49	383	13973	35	130
	MOT17-12-FRCNN	47.3	58.0	18	47	37	4521	11	84
	MOT17-14-FRCNN	36.2	49.0	10	44	629	11061	108	359
	MOT17-01-SDP	47.8	57.8	38	42	346	3008	10	95
	MOT17-03-SDP	78.2	77.3	62	9	3778	18879	132	16219
	MOT17-06-SDP	60.3	65.1	30	29	305	4345	33	144
	MOT17-07-SDP	45.8	55.0	13	30	285	8793	71	483
	MOT17-08-SDP	34.8	47.7	24	45	429	13288	48	202
	MOT17-12-SDP	47.3	60.7	20	46	158	4394	14	85
	MOT17-14-SDP	38.3	51.4	9	42	630	10662	109	376
MOT16	MOT16-01	48.3	58.2	35	43	78	3217	10	38
	MOT16-03	73.0	69.9	54	11	3732	24329	159	24311
	MOT16-06	58.2	64.7	28	35	249	4548	29	113
	MOT16-07	45.6	53.4	13	30	189	8637	57	297
	MOT16-08	43.4	55.7	29	38	284	9149	32	146
	MOT16-12	50.2	64.0	21	43	44	4072	11	68
	MOT16-14	36.1	48.8	7	47	268	11449	91	323
2D MOT15	ADL-Rundle-1	39.6	60.8	41	6	2277	3303	44	325
	ADL-Rundle-3	59.2	69.9	52	16	902	3217	29	153
	AVG-TownCentre	61.8	67.3	43	15	417	2217	99	20
	ETH-Crossing	57.6	69.3	27	35	35	387	3	2
	ETH-Jelmoli	51.4	67.1	40	31	520	701	12	20
	ETH-Linthescher	53.7	32	21	98	318	3795	21	11
	KITTI-16	36.2	32.7	29	6	456	521	108	57
	KITTI-19	43.3	49.4	18	27	467	2315	249	135
	PETS09-S2L2	56.9	43.6	21	5	476	3531	152	180
	TUD-Crossing	88.0	90.9	85	0	64	62	6	13
	Venice-1	45.8	62.1	53	18	905	1561	7	30

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

5 Conclusions

This work addresses the multiple object tracking problem (MOT): Using sensor data, all objects of a common object class are to be localized and, additionally, their motion history is to be captured in the form of trajectories. The automated acquisition of movements as a numerical representation has numerous applications, such as in autonomous driving, monitoring of safety-relevant areas, and behavior analysis. Therefore, multiple object tracking has been a focus of research for a long time and numerous methods to tackle the problem have been developed. This work focuses on MOT based on video data since video sensors are cheap, widely available, and captured images contain rich information that helps to tackle the difficulties of MOT.

The commonly used approach of computing object detections in each frame of a video recording leads to a considerable reduction of the computational costs during the assignment of objects to the image content. It thus simplifies the computation of trajectories. However, this implies that the accuracy of such tracking methods depends substantially on the input detections which is disadvantageous for several reasons: (i) Errors generated by a detector significantly affect the tracking result. (ii) Using image features only between detections can be unstable as these pairwise features commonly rely on certain appearance and motion assumptions between detections that are not always satisfied. (iii) Ensuring temporal consistency within the data association is difficult due to the limited remaining information.

In this thesis we demonstrate that by addressing the aforementioned limitations of current tracking-by-detection methods, significant improvements can be achieved. To this end, in the first approach of this thesis, a global fusion formulation is presented that allows object detections to be combined with additional signals so that the dependence on object detections and on the accuracy of features extracted from the object detections is reduced. A novel approximate solver adapted to the corresponding optimization problem is presented that is crucial to obtain near-optimal solutions. The second main approach presented in this thesis addresses a comprehensive integration of the spatio-temporal information provided by object detections. A novel MOT formulation is presented that integrates path connectivity priors and higher-order consistencies. A global optimal solver for the problem is proposed, making it possible, to our knowledge for the first time, to compute trajectories from a tracking model with path connectivity priors and higher-order consistencies without relying on a heuristic solver. We thus argue that all existing MOT methods do not satisfactorily exploit the available information.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

HO-MOT with Signal Fusion

The first part of this thesis addresses the severe dependence of the tracking-by-detection paradigm on computed detections. To this end, object detections are fused with additional signals using a global optimization formulation. The fusion result is obtained from a weighted graph labeling problem in which each input signal is assigned the corresponding object. Consistency is ensured within each signal and between different signal types. This allows compensating errors occurring in an individual signal, *e.g.*, missing object detections, in contrast to tracking-by-detection based approaches.

In the proposed formulation, all pairwise connections assigned to an object contribute to the evaluation of a trajectory. Accordingly, it corresponds to an implicit HO-MOT method. The underlying optimization problem is formulated as a BQP. Since it is $\mathcal{NP}$-hard, the computation of a solution is difficult. A solver is presented that operates in the continuous domain by applying the Frank-Wolfe method to the continuous relaxation. However, a direct application of this solver does not provide a sufficiently good approximation. Several improvements are proposed that result in the solver converging faster on the one hand and providing better approximations on the other hand. In the evaluations of the most difficult sequence of the MOT16 dataset, the solver provides almost optimal solutions. Moreover, conducted comparisons with a global optimal BQP solver reveal that the optimal solver delivers a result after 1000 seconds that is considerably worse in terms of objective value and tracking accuracy than the result of the proposed approximate solver which terminates after 28 seconds. A vanilla Frank-Wolfe implementation yields a MOTA score of 14.2%. The optimal solver produces a MOTA score of 24.9% in the given time. The proposed solver improves considerably upon these results, achieving 27.5% MOTA. In order to separate the influence of erroneous detections from the quality of the solver, the graph labeling problem is also evaluated on all sequences of the MOT16 training set using ground truth people detections. Conducted experiments show that the proposed solver improvements lead to significantly better solutions on all sequences with respect to the objective value as well as tracking metrics. For instance, the proposed solver achieves 100% IDF1 on the MOT16-09 sequence. Without using all proposed solver improvements, a significantly worse solution with an IDF1 value of 80.6% is achieved. Thus, the proposed solver is crucial for the application of the fusion formulation in practice. The benefits of fusion are demonstrated by combining head detections and people detections. Heads can often be detected even if persons are partially occluded. They also facilitate the detection of false positive detections that do not go with head detections. Accordingly, the fusion leads to significantly better tracking results. The method, which we call *Frank-Wolfe Tracker* (FWT), evaluated on the MOT16 training data, achieves a reduction of false positives by more than 50% and an improvement of the MOTA value by 15%. On the MOT16 and MOT17 test data, FWT delivered state-of-the-art results at the time the work was published and was able to win the MOT 2017 tracking challenge of the CVPR 2017.

The advantage of signal fusion is also demonstrated by a second method in which video data is combined with IMU data. Here, it is assumed that each person to be tracked wears an IMU sensor on his or her back and that the camera is static and calibrated. We refer to the task of tracking multiple people in a video and simultaneously assigning them

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

to IMU sensors as *Video Inertial Multiple People Tracking* (VIMPT). This task poses a challenging problem. It is not sufficient to generate trajectories from IMU data and then map them to image data since the necessary double integration of the acceleration signal is numerically unstable and initial states are unknown. Instead, 3D orientations as well as accelerations must be assigned to the 2D image information, which is again a difficult task on its own since a 2D projection creates ambiguities. In addition, there are often several persons for whom the orientations as well as accelerations measured at one point in time are similar.

To resolve these ambiguities, the problem is considered as a weighted graph labeling problem in which detections are assigned to IMU devices. Therefore, image data must be linked with the IMU signals, which is mainly performed using two features. First, the orientation of a person is estimated from the image data of a detection box and compared to the measurements of the IMUs. For this purpose, a neural network is presented which corrects perspective effects that depend on the relative position of a person to the camera. On the test data of the new recorded VIMPT2019 dataset, the orientation estimation is within an error tolerance of 45° and 30° in 96.2% and 88.1% of the estimated cases, respectively. Second, measured accelerations and initial velocities approximated in the image are used to ensure consistency between assignments of video content and IMU signals. By embedding these features into the global fusion formulation, robust trajectories are generated.

The resulting method *Video Inertial Tracker* (VIT) addresses the aforementioned problems of the tracking-by-detection paradigm: By integrating actual measured motion information, VIT is less dependent on predefined motion and appearance models. The problem of error-prone features extracted from object detections is thus reduced.

This becomes evident in the soccer sequences of the VIMPT2019 dataset. Due to the characteristics of the recordings (especially abrupt movements and similar appearance due to soccer jerseys), the implicit assumptions of video-based tracking methods are violated, reducing the tracking accuracy significantly, so that the compared video-based MOT methods reach no more than 45.8% IDF1 in the experiments. In contrast, VIT yields a very high IDF1 value of 91.8%. In addition, VIT significantly reduces the dependence on object detections through a proposed reconstruction method that recovers missing detections using the computed trajectories in the video together with the corresponding IMU data. For example, the IDF1 metric on the VIMPT2019 test data decreases from 91% to 81% when 20% of the detections are (synthetically) removed. In contrast, the IDF1 value drops dramatically from 78% to 43% when video-based linear interpolation is employed. Thus, VIT addresses the problems of the tracking-by-detection paradigm, which exhibits a high dependence on the input detections. Furthermore, VIT allows identifying for each IMU signal the corresponding person in the image independent of the outward appearance and across sequences. The experiments show that the trajectory-to-person assignment accuracy of VIT on the VIMPT2019 test data is 100%.

Thus, VIT allows identifying persons in a video and tracking them robustly at the same time. Weaknesses of the tracking-by-detection paradigm (in particular, missing detections and erroneous edge weights) are effectively addressed by the fusion with IMU data. We consider the VIMPT task as a useful complement to the MOT problem provided that the setup is applicable. VIT automatically identifies people in videos,

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

maps them to IMU signals, and simultaneously computes all people trajectories using an MOT formulation.

In summary, the fusion of object detections with additional signals, both additional image content and additional sensors, effectively addresses disadvantages of the tracking-by-detection paradigm. By using complementary signals, weaknesses of individual signals can be compensated and advantages combined. The proposed global fusion formulation using the novel solver achieves significant improvements over the traditional approach of using only object detections. By exploiting more of the available information, robustness to erroneous input is greatly improved.

Lifted Disjoint Paths

The second part of this thesis addresses the limited exploitation of provided information by tracking-by-detection methods. To this end, a novel data association model, called *Lifted Disjoint Paths* (LDP), is introduced which extends the disjoint paths problem by lifted edges. The model becomes more expressive but retains the natural representation of trajectories of the disjoint paths formulation.

Ensuring temporal consistency by extracting image features only from object detections is a strong limitation, challenging, and error-prone. This is particularly severe as existing MOT methods do not fully exploit the information provided by the object detections. Some of the previous works use a simple assignment model, which can be solved to global optimality, but do not exploit long-range interactions. Other works use complex HO-MOT models but then rely on heuristic solvers. This has decisive disadvantages: Many such methods depend on an initial solution, which can have a negative impact on the robustness if not chosen properly for each problem instance. Most importantly, a heuristic solver usually does not provide an optimal solution. Therefore, despite many decades of research in MOT, input information for existing methods is exploited only to a limited extent. As a result, misleading weight costs, false positive as well as false negative detections cannot be properly recognized and corrected which deteriorates the accuracy of trajectories.

In contrast, the proposed LDP tracker incorporates, to our knowledge, the first global optimization method that integrates long-range interactions. Since no errors are generated in the process, improvements in the input data (better detection quality and more accurate features) lead directly to improvements in tracking accuracy. The proposed optimal solver operates on a BLP formulation of the LDP problem. Several linear relaxations of the BLP are derived and integrated into efficient separation algorithms. Optimal solutions are obtained from a cutting-plane algorithm. The integration of lifted edges in LDP allows integrating path connectivity priors. Consequently, long-term temporal information is robustly integrated and long-range interactions are exploited.

To demonstrate the advantage of LDP, we propose to fuse different features in a neural network producing weight costs for the data association problem. Employed features can be categorized into (i) spatio-temporal cues, (ii) appearance cues (based on DeepMatching and a re-identification method), and (iii) novel global context normalization features that evaluate the discriminative power of pairwise costs within a global context

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

between frames. The fusion of all these cues results in accurate and robust estimates of whether two detections belong to the same person up to a time interval of 2 seconds.

Conducted experiments show the advantage of LDP. The assignment quality is analyzed on the MOT17 training dataset. When using edges that connect detections with a maximum time difference of 0.3 seconds, the IDF1 metric yields 55.7%. Using all edges up to a time interval of 2 seconds, the metric is significantly improved to 64.3%, demonstrating that LDP can effectively exploit long-term temporal information and long-range interactions. Furthermore, the data association is near-optimal, as a maximum IDF1 of 66.8% is possible in the experiment with the given detections.

Finally, the evaluations on the test data of MOT15/16/17 show that LDP has significantly improved the state of the art, when the work was published. On MOT15, MOT16, and MOT17, the improvement in IDF1 relative to the best competing method is 13.3, 12.2, and 13.3 percentage points, respectively.

Overall, LDP produces significantly more consistent trajectories than competing methods thanks to the integration of long-range interactions. Crucial for LDP is the novel global solver that, to our knowledge, is the first to provide assignments for an implicit HO-MOT model with long-range interactions without a heuristic solver. Analyses also show that LDP leads to near-optimal data association results. Major improvements for tracking-by-detection methods on the used benchmarks can therefore no longer be expected from an improved data association but from the use of better detections as well as improved methods that reconstruct missing detections within a trajectory.

Future Work

In this work, we present higher-order MOT methods that improve the tracking accuracy of existing approaches by better exploiting available information. Nevertheless, the proposed methods are subject to some limitations that can be addressed in future work.

We recall that the number of objects that appear in a recording must be upper bounded in the graph labeling formulation of Chapter 3. For the tracker FWT, finding a sufficiently large bound could be automated, *e.g.*, by checking whether increasing the bound leads to an improved objective value. Note, however, that due to the hierarchical solving scheme, the upper bound only needs to be roughly estimated. Alternatively, the LDP method of Chapter 4 performs MOT without the need of knowing a priori the number of objects to be tracked. In contrast, the VIMPT setup assumes that each person to be tracked wears an IMU sensor. Thus, VIMPT provides by its setup the upper bound and the number of persons to be tracked. Furthermore, VIT could be extended to additionally track people in a video that are not equipped with an IMU sensor. Since VIT allows for the identification of a trajectory that matches an IMU sensor, these trajectories could be removed so that a purely video-based MOT method could be applied on the remaining detections.

The combination of head detections with person detections in FWT allows compensating errors caused by the object detector or weight costs. However, conducted evaluations reveal that relying essentially on DeepMatching is subject to identity switches, as DeepMatching does not differentiate if the image content within a detection belongs

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

to the background or another object. Accordingly, the fusion framework benefits from more advanced image features around a head to make re-identification more stable. Also, improvements could be achieved by extracting a foreground/background mask for each person detection [263] or in a post-processing step where tracklet consistency is checked. In addition, for recordings taken from low altitude, the head of a partially occluded person is more likely to be missed by the head detector so that the advantage of the fusion diminishes. This limitation could be compensated by augmenting people detections with other signals, *e.g.*, joint detections [101] or segmentation masks [190]. Moreover, since the global fusion formulation is independent of the number of signals to be fused, the method could be further improved by fusing even more input signals at once.

The accuracy of the proposed MOT methods of this thesis can be further improved by incorporating edges connecting detections over even longer temporal distances or using multiple signal types. However, constructing corresponding edges weights is complicated. This could be addressed by transforming the proposed approaches into end-to-end learnable MOT methods that are directly trained on tracking metrics, inspired by the current progress in this direction [104, 264] for other tracking approaches. To this end, a differentiable Frank-Wolfe solver [265] could be used for the methods of Chapter 3, and a method to perform differentiation of combinatorial solvers [266] for the LDP method of Chapter 4.

VIT shows some limitations with respect to practicability as cameras need to be calibrated. This could be avoided if the comparisons between IMU measurements and corresponding estimations from image data were not made independently frame by frame but directly on entire tracklets. This avoids the registration of coordinate systems as it allows to reason from relative motions, similar to Sun *et al.* [267]. Also, attention mechanisms, *e.g.*, in the spirit of Xu *et al.* [82], could be used to perform orientation estimation only for those detections of a tracklet that do not contain impaired visual information, such as partial occlusions or motion blur.

Evaluations of VIT show that a fusion of video data with IMU signals improves MPT methods considerably. We note that the same concept could be applied to track other objects which would extend the setup to *Video Inertial Multiple Object Tracking* (VIMOT). This is especially interesting for cases in which it is very hard to distinguish individual objects. One application in mind is tracking and identifying individual animals in a herd.

The extension of the disjoint paths problem by lifted edges in Chapter 4 leads to an accurate HO-MOT method. Base and lifted edges are constructed in the same way (with respect to costs and temporal distances). However, the formulation also allows treating the two edge types differently, for example, by distinguishing between short edges in the base graph and long edges with uncertain costs in the lifted graph. Future research can therefore look more closely at the interaction between the two edge types, which could lead to even better tracking results.

The LDP method is the first to solve HO-MOT problem instances globally. Yet, it is limited by size of the problem instances. To make HO-MOT scalable to even larger instances, finding tighter relaxations for the LDP problem could decrease runtime as well as an approximate LDP solver, *e.g.*, based on dual decomposition [123, 124, 268],

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

to quickly generate upper bounds.

Regarding the optimization problems proposed in this thesis, it should be noted that the corresponding solver from Chapter 3 for the weighted graph labeling problem and the solver from Chapter 4 for the lifted disjoint paths problem are not inherently tied to MOT. For instance, the Frank-Wolfe solver could potentially be used for multi-camera MOT [269] or to compute superpixels [270]. Additionally, the LDP problem poses a new $\mathcal{NP}$-hard decision problem that could get attention in computational complexity theory. We anticipate that these solvers will stimulate new research in various areas of computer vision and beyond.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

Bibliography

[1] Markus Maurer, J Christian Gerdes, Barbara Lenz, and Hermann Winner. *Autonomous driving: technical, legal and social aspects*. Springer Nature, 2016.

[2] Andreas Geiger, Philip Lenz, and Raquel Urtasun. “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[3] Davi Frossard, Eric Kee, and Raquel Urtasun. “DeepSignals: Predicting Intent of Drivers Through Visual Signals”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2019.

[4] Wenjie Luo, Bin Yang, and Raquel Urtasun. “Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[5] Sarthak Sharma, Junaid Ahmed Ansari, J Krishna Murthy, and K Madhava Krishna. “Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking”. In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2018.

[6] Gültekin Gündüz and Tankut Acarman. “A lightweight online multiple object vehicle tracking method”. In: *IEEE Intelligent Vehicles Symposium (IV)*. 2018.

[7] Bing He, Jia Li, Yifan Zhao, and Yonghong Tian. “Part-regularized near-duplicate vehicle re-identification”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[8] Giorgos Bouritsas, Stelios Daveas, Antonios Danelakis, and Stelios CA Thomopoulos. “Automated Real-time Anomaly Detection in Human Trajectories using Sequence to Sequence Networks”. In: *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2019.

[9] Yan Xu, Xi Ouyang, Yu Cheng, Shining Yu, Lin Xiong, Choon-Ching Ng, Sugiri Pranata, Shengmei Shen, and Junliang Xing. “Dual-mode vehicle motion pattern learning for high performance road traffic anomaly detection”. In: *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[10] Milind Naphade, Zheng Tang, Ming-Ching Chang, David C. Anastasiu, Anuj Sharma, Rama Chellappa, Shuo Wang, Pranamesh Chakraborty, Tingting Huang, Jenq-Neng Hwang, and Siwei Lyu. “The 2019 AI City Challenge”. In: *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019.

[11] Marco Cristani, Alessio Del Bue, Vittorio Murino, Francesco Setti, and Alessandro Vinciarelli. *The Visual Social Distancing Problem*. 2020. arXiv: `2005.04813 [cs.CV]`.

[12] Andrew J Newman, KC Daniel, and David P Oulton. “New insights into retail space and format planning from customer-tracking data”. In: *Journal of Retailing and Consumer Services* 9.5 (2002), pp. 253–258.

[13] Alexandre Alahi, Judson Wilson, Li Fei-Fei, and Silvio Savarese. “Unsupervised camera localization in crowded spaces”. In: *Proceedings of the IEEE Conference on Robotics and Automation (ICRA)*. 2017.

[14] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. “Socially-aware large-scale crowd forecasting”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[15] Colby T Jeffries. *Sports Analytics With Computer Vision*. The College of Wooster, 2018. URL: `https://openworks.wooster.edu/independentstudy/8103`.

[16] Alexandre Alahi, Yannick Boursier, Laurent Jacques, and Pierre Vandergheynst. “Sport players detection and tracking with a mixed network of planar and omnidirectional cameras”. In: *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*. 2009.

[17] Wei-Lwun Lu, Jo-Anne Ting, James J Little, and Kevin P Murphy. “Learning to track and identify players from broadcast sports videos”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 35.7 (2013), pp. 1704–1716.

[18] Runze Li and Bir Bhanu. “Fine-Grained Visual Dribbling Style Analysis for Soccer Videos With Augmented Dribble Energy Image”. In: *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019.

[19] Rajkumar Theagarajan, Federico Pala, Xiu Zhang, and Bir Bhanu. “Soccer: Who has the ball? Generating visual analytics and player statistics”. In: *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018.

[20] Margarete Boos, Johannes Pritz, Simon Lange, and Michael Belz. “Leadership in moving human groups”. In: *PLoS Comput Biol* 10.4 (2014).

[21] Dirk Helbing and Peter Molnar. “Social force model for pedestrian dynamics”. In: *Physical review E* 51.5 (1995).

[22] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. “You’ll never walk alone: Modeling social behavior for multi-target tracking”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2009.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[23] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn. “Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2011.

[24] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. “Social gan: Socially acceptable trajectories with generative adversarial networks”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[25] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. “Social LSTM: Human Trajectory Prediction in Crowded Spaces”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[26] Alexandre Alahi, Vignesh Ramanathan, Kratarth Goel, Alexandre Robicquet, Amir A Sadeghian, Li Fei-Fei, and Silvio Savarese. “Learning to Predict Human Behavior in Crowded Scenes”. In: *Group and Crowd Behavior for Computer Vision*. Elsevier, 2017, pp. 183–207.

[27] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. “Learning social etiquette: Human trajectory understanding in crowded scenes”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2016.

[28] Shuai Yi, Hongsheng Li, and Xiaogang Wang. “Understanding Pedestrian Behaviors From Stationary Crowd Groups”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[29] Willem Bouten, Edwin W Baaij, Judy Shamoun-Baranes, and Kees CJ Camphuysen. “A flexible GPS tracking system for studying bird behaviour at multiple scales”. In: *Journal of Ornithology* 154.2 (2013), pp. 571–580.

[30] Alfonso Pérez-Escudero, Julián Vicente-Page, Robert C Hinz, Sara Arganda, and Gonzalo G De Polavieja. “idTracker: tracking individuals in a group by automatic identification of unmarked animals”. In: *Nature methods* 11.7 (2014), pp. 743–748.

[31] Franziska Boenisch, Benjamin Rosemann, Benjamin Wild, David Dormagen, Fernando Wario, and Tim Landgraf. “Tracking all members of a honey bee colony over their lifetime using learned models of correspondence”. In: *Frontiers in Robotics and AI* 5 (2018), p. 35.

[32] Katarzyna Bozek, Laetitia Hebert, Alexander S Mikheyev, and Greg J Stephens. “Towards dense object tracking in a 2D honeybee hive”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[33] Joe R Riley, Uwe Greggers, Alan D Smith, Don R Reynolds, and Randolf Menzel. “The flight paths of honeybees recruited by the waggle dance”. In: *Nature* 435.7039 (2005), pp. 205–207.

[34] Gaoang Wang, Jenq-Neng Hwang, Kresimir Williams, and George Cutter. “Closed-loop tracking-by-detection for ROV-based multiple fish tracking”. In: *ICPR Workshop on Computer Vision for Analysis of Underwater Imagery (CVAUI)*. IEEE. 2016.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[35] Malte Pedersen, Joakim Bruslund Haurum, Stefan Hein Bengtson, and Thomas B Moeslund. “3D-ZeF: A 3D Zebrafish Tracking Benchmark Dataset”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[36] Zia Khan, Tucker Balch, and Frank Dellaert. “An MCMC-based particle filter for tracking multiple interacting targets”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2004.

[37] Margrit Betke, Diane E Hirsh, Angshuman Bagchi, Nickolay I Hristov, Nicholas C Makris, and Thomas H Kunz. “Tracking large variable numbers of objects in clutter”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2007.

[38] Laura Leal Taixé, Matthias Heydt, Axel Rosenhahn, and Bodo Rosenhahn. “Automatic tracking of swimming microorganisms in 4D digital in-line holography data”. In: *Workshop on Motion and Video Computing (WMVC)*. IEEE. 2009.

[39] Laura Leal-Taixé, Matthias Heydt, Sebastian Weiße, Axel Rosenhahn, and Bodo Rosenhahn. “Classification of swimming microorganisms motion patterns in 4D digital in-line holography data”. In: *Joint Pattern Recognition Symposium*. Springer. 2010, pp. 283–292.

[40] Markus Rempfler, Jan-Hendrik Lange, Florian Jug, Corinna Blasse, Eugene W Myers, Bjoern H Menze, and Bjoern Andres. “Efficient algorithms for moral lineage tracing”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[41] Florian Jug, Evgeny Levinkov, Corinna Blasse, Eugene W Myers, and Bjoern Andres. “Moral lineage tracing”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[42] Junya Hayashida, Kazuya Nishimura, and Ryoma Bise. “MPM: joint representation of motion and position map for cell tracking”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[43] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. *MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking*. 2015. arXiv: `1504.01942 [cs.CV]`.

[44] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. *MOT16: A Benchmark for Multi-Object Tracking*. 2016. arXiv: `1603.00831 [cs.CV]`.

[45] Ankith Manjunath, Ying Liu, Bernardo Henriques, and Armin Engstle. “Radar based object detection and tracking for autonomous driving”. In: *IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*. 2018.

[46] ITUNews. *The price of LiDAR is falling. Will driverless cars be on the road sooner?* [Online; accessed June, 2021]. URL: `https://news.itu.int/the-price-of-lidar-is-falling/`.

[47] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. “Object detection with discriminatively trained part-based models”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 32.9 (2010), pp. 1627–1645.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252.

[49] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. "Deep learning for generic object detection: A survey". In: *International Journal of Computer Vision (IJCV)* 128.2 (2020), pp. 261–318.

[50] **Andrea Hornakova**, **Roberto Henschel**, Bodo Rosenhahn, and Paul Swoboda. "Lifted Disjoint Paths with Application in Multiple Object Tracking". In: *International Conference on Machine Learning (ICML)*. 2020. *The highlighted authors share first authorship with equal contributions.*

[51] Evgeny Levinkov, Jonas Uhrig, Siyu Tang, Mohamed Omran, Eldar Insafutdinov, Alexander Kirillov, Carsten Rother, Thomas Brox, Bernt Schiele, and Bjoern Andres. "Joint graph decomposition & node labeling: Problem, algorithms, applications". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[52] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[53] William Brendel, Mohamed Amer, and Sinisa Todorovic. "Multiobject tracking as maximum weight independent set". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

[54] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. "Subgraph decomposition for multi-target tracking". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[55] Roberto Henschel, Laura Leal-Taixé, Daniel Cremers, and Bodo Rosenhahn. "Fusion of Head and Full-Body Detectors for Multi-Object Tracking". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018.

[56] Zheng Wu, Ashwin Thangali, Stan Sclaroff, and Margrit Betke. "Coupling detection and data association for multiple object tracking". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2012.

[57] Bastian Leibe, Konrad Schindler, and Luc Van Gool. "Coupled detection and trajectory estimation for multi-object tracking". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2007.

[58] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. "Tracking without bells and whistles". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[59] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. "Towards real-time multi-object tracking". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[60] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. “Tracking objects as points”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2020.

[61] Li Zhang, Yuan Li, and Ramakant Nevatia. “Global data association for multi-object tracking using network flows”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2008.

[62] Laura Leal-Taixé, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, and Stefan Roth. *Tracking the trackers: an analysis of the state of the art in multiple object tracking*. 2017. arXiv: `1704.02781 [cs.CV]`.

[63] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. “Multi-person tracking by multicut and deep matching”. In: *ECCV Workshop on Benchmarking Multi-Target Tracking (ECCVW)*. Springer. 2016.

[64] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. “Multiple people tracking by lifted multicut and person reidentification”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[65] Wongun Choi. “Near-online multi-target tracking with aggregated local flow descriptor”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[66] Jeany Son, Mooyeol Baek, Minsu Cho, and Bohyung Han. “Multi-Object Tracking with Quadruplet Convolutional Neural Networks”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[67] Jerome Berclaz, Francois Fleuret, Engin Turetken, and Pascal Fua. “Multiple object tracking using k-shortest paths optimization”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 33.9 (2011), pp. 1806–1819.

[68] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, Xiaowei Zhao, and Tae-Kyun Kim. *Multiple object tracking: A literature review*. 2014. arXiv: `1409.7618 [cs.CV]`.

[69] Hans-Otto Georgii. *Stochastics: introduction to probability and statistics*. Walter de Gruyter, 2012.

[70] Amir Roshan Zamir, Afshin Dehghan, and Mubarak Shah. “GMCP-tracker: Global multi-object tracking using generalized minimum clique graphs”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2012.

[71] Donald Reid. “An algorithm for tracking multiple targets”. In: *IEEE transactions on Automatic Control* 24.6 (1979), pp. 843–854.

[72] Charles Morefield. “Application of 0-1 integer programming to multitarget tracking problems”. In: *IEEE Transactions on Automatic Control* 22.3 (1977), pp. 302–312.

[73] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2015.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[74] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. "LIBLINEAR: A Library for Large Linear Classification". In: *Journal of Machine Learning Research (JMLR)* 9 (2008), pp. 1871–1874.

[75] Roberto Henschel, Timo von Marcard, and Bodo Rosenhahn. *VIMPT2019 - Video Inertial Multiple People Tracking dataset.* [Online, accessed 12. Dezember 2020]. 2019. URL: `https://www.tnt.uni-hannover.de/de/project/VIMPT2019/`.

[76] Cheng-Hao Kuo and Ram Nevatia. "How does person identity recognition help multi-person tracking?" In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

[77] Margret Keuper, Siyu Tang, Bjoern Andres, Thomas Brox, and Bernt Schiele. "Motion Segmentation & Multiple Object Tracking by Correlation Co-Clustering". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 42.1 (2020), pp. 140–153.

[78] Ergys Ristani and Carlo Tomasi. "Features for Multi-Target Multi-Camera Tracking and Re-Identification". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[79] Chanho Kim, Fuxin Li, and James M Rehg. "Multi-object tracking with neural gating using bilinear lstm". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2018.

[80] Ji Zhu, Hua Yang, Nian Liu, Minyoung Kim, Wenjun Zhang, and Ming-Hsuan Yang. "Online Multi-Object Tracking with Dual Matching Attention Networks". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2018.

[81] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu. "Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[82] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. "Spatial-temporal relation networks for multi-object tracking". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[83] Xingping Dong, Jianbing Shen, Wenguan Wang, Ling Shao, Haibin Ling, and Fatih Porikli. "Dynamical Hyperparameter Optimization via Deep Reinforcement Learning in Tracking". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 43.5 (2021), pp. 1515–1529.

[84] Zhiyuan Liang and Jianbing Shen. "Local Semantic Siamese Networks for Fast Tracking". In: *IEEE Transactions on Image Processing (TIP)* 29 (2020), pp. 3351–3364.

[85] Aubrey B Poore. "Multidimensional assignment problems arising in multitarget and multisensor tracking". In: *Nonlinear Assignment Problems* (2000), pp. 13–38.

[86] Michael R Garey and David S Johnson. *Computers and intractability.* Vol. 174. freeman San Francisco, 1979.

[87] Francesco Solera, Simone Calderara, and Rita Cucchiara. "Learning to divide and conquer for online multi-target tracking". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[88] Peng Chu and Haibin Ling. “Famnet: Joint learning of feature, affinity and multidimensional assignment for online multiple object tracking”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019.

[89] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. “Simple online and realtime tracking”. In: *Proceedings of the IEEE Conference on Image Processing (ICIP)*. 2016.

[90] Erik Bochinski, Volker Eiselein, and Thomas Sikora. “High-speed tracking-by-detection without using image information”. In: *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 2017.

[91] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. “Tracking the untrackable: Learning to track multiple cues with long-term dependencies”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[92] Michael D Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. “Online multiperson tracking-by-detection from a single, uncalibrated camera”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 33.9 (2010), pp. 1820–1833.

[93] Bo Wu and Ram Nevatia. “Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors”. In: *International Journal of Computer Vision (IJCV)* 75.2 (2007), pp. 247–266.

[94] Cor J Veenman, Marcel JT Reinders, and Eric Backer. “Resolving motion correspondence for densely moving points”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 23.1 (2001), pp. 54–72.

[95] Bo Yang and Ram Nevatia. “Multi-target tracking by online learning of non-linear motion patterns and robust appearance models”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[96] Rudolph Emil Kalman. “A new approach to linear filtering and prediction problems”. In: *Journal of Basic Engineering* 82.1 (1960), pp. 35–45.

[97] Nicolai Wojke and Alex Bewley. “Deep Cosine Metric Learning for Person Re-identification”. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018.

[98] Zia Khan, Tucker Balch, and Frank Dellaert. “MCMC-based particle filtering for tracking a variable number of interacting targets”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 27.11 (2005).

[99] Pierre Del Moral. “Nonlinear filtering: Interacting particle resolution”. In: *Comptes Rendus de l’Académie des Sciences-Series I-Mathematics* 325.6 (1997), pp. 653–658.

[100] Harold W Kuhn. “The Hungarian method for the assignment problem”. In: *Naval research logistics quarterly* 2.1-2 (1955), pp. 83–97.

[101] Roberto Henschel, Yunzhe Zou, and Bodo Rosenhahn. “Multiple people tracking using body and joint detections”. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[102] Hamed Pirsiavash, Deva Ramanan, and Charless C Fowlkes. “Globally-optimal greedy algorithms for tracking a variable number of objects”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

[103] Visesh Chari, Simon Lacoste-Julien, Ivan Laptev, and Josef Sivic. “On pairwise costs for network flow multi-object tracking”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[104] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. “Deep network flow for multi-object tracking”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[105] Guillem Brasó and Laura Leal-Taixé. “Learning a Neural Solver for Multiple Object Tracking”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

[106] Roberto Henschel, Laura Leal-Taixé, and Bodo Rosenhahn. “Efficient multiple people tracking using minimum cost arborescences”. In: *German Conference on Pattern Recognition (GCPR)*. 2014.

[107] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. *Network flows*. Cambridge, Mass.: Alfred P. Sloan School of Management, Massachusetts, 1988.

[108] Congchao Wang, Yizhi Wang, Yinxue Wang, Chiung-Ting Wu, and Guoqiang Yu. “muSSP: Efficient Min-cost Flow Algorithm for Multi-object Tracking”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2019.

[109] Longyin Wen, Dawei Du, Shengkun Li, Xiao Bian, and Siwei Lyu. “Learning non-uniform hypergraph for multi-object tracking”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019.

[110] Martin Hofmann, Daniel Wolf, and Gerhard Rigoll. “Hypergraphs for joint multi-view reconstruction and multi-object tracking”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013.

[111] Asad A Butt and Robert T Collins. “Multiple target tracking using frame triplets”. In: *Asian Conference on Computer Vision*. Springer. 2012.

[112] Chetan Arora and Amir Globerson. “Higher order matching for consistent multiple target tracking”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2013.

[113] Roy L Streit and Tod E Luginbuhl. *Probabilistic multi-hypothesis tracking*. Tech. rep. NAVAL UNDERWATER SYSTEMS CENTER NEWPORT RI, 1995.

[114] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. “Multiple hypothesis tracking revisited”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[115] Corinne Feremans, Martine Labbé, and Gilbert Laporte. “Generalized network design problems”. In: *European Journal of Operational Research* 148.1 (2003), pp. 1–13.

[116] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. “Correlation clustering”. In: *Machine learning* 56.1-3 (2004), pp. 89–113.

[117] Marguerite Frank and Philip Wolfe. “An algorithm for quadratic programming”. In: *Naval Research Logistics Quarterly* (1956).

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[118] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. “Recovering accurate 3d human pose in the wild using imus and a moving camera”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

[119] I-Lin Wang. “Multicommodity network flows: A survey, Part I: Applications and Formulations”. In: *International Journal of Operations Research* 15.4 (2018).

[120] Roberto Henschel, Laura Leal-Taixé, and Bodo Rosenhahn. “Solving multiple people tracking in a minimum cost arborescence”. In: *IEEE Winter Applications and Computer Vision Workshops (WACVW)*. 2015.

[121] Roberto Henschel, Timo von Marcard, and Bodo Rosenhahn. “Simultaneous Identification and Tracking of Multiple People using Video and IMUs”. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019.

[122] Roberto Henschel, Timo von Marcard, and Bodo Rosenhahn. “Accurate Long-Term Multiple People Tracking using Video and Body-Worn IMUs”. In: *Transactions on Image Processing (TIP)* 29 (2020), pp. 8476–8489.

[123] Andrea Hornakova, Timo Kaiser, Bodo Rosenhahn, Paul Swoboda, and Roberto Henschel. “Higher Order Multiple Object Tracking for Crowded Scenes”. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2021.

[124] Andrea Hornakova, Timo Kaiser, Paul Swoboda, Michal Rolinek, Bodo Rosenhahn, and Roberto Henschel. “Making Higher Order MOT Scalable: An Efficient Approximate Solver for Lifted Disjoint Paths”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2021.

[125] Dieter Jungnickel. *Graphs, networks and algorithms*. Springer, 2005.

[126] Edsger W Dijkstra. “A note on two problems in connexion with graphs”. In: *Numerische mathematik* 1.1 (1959), pp. 269–271.

[127] Tom M Mitchell et al. “Machine learning. 1997”. In: *Burr Ridge, IL: McGraw Hill* 45.37 (1997), pp. 870–877.

[128] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[129] John Ashworth Nelder and Robert WM Wedderburn. “Generalized linear models”. In: *Journal of the Royal Statistical Society: Series A (General)* 135.3 (1972), pp. 370–384.

[130] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009.

[131] Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik, Luigi Troiano, Roberto Tagliaferri, and Francisco Herrera. “Deep learning in video multi-object tracking: A survey”. In: *Neurocomputing* 381 (2020), pp. 61–88.

[132] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. *An analysis of deep neural network models for practical applications*. 2016. arXiv: `1605.07678 [cs.CV]`.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[133] Maithra Raghu and Eric Schmidt. *A survey of deep learning for scientific discovery*. 2020. arXiv: `2003.11755 [cs.LG]`.

[134] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[135] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[136] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2014.

[137] Yann LeCun, Yoshua Bengio, et al. "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10 (1995).

[138] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[139] Allan Pinkus. "Approximation theory of the MLP model in neural networks". In: *Acta numerica* 8.1 (1999), pp. 143–195.

[140] Patrick Kidger and Terry Lyons. "Universal approximation with deep narrow networks". In: *Conference on Learning Theory*. 2020.

[141] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[142] Michael Ulbrich and Stefan Ulbrich. *Nichtlineare Optimierung*. Springer-Verlag, 2012.

[143] Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *The annals of mathematical statistics* (1951), pp. 400–407.

[144] H Martin Bücker, George Corliss, Paul Hovland, Uwe Naumann, and Boyana Norris. *Automatic differentiation: applications, theory, and implementations*. Vol. 50. Springer Science & Business Media, 2006.

[145] Paul J Werbos. "Applications of advances in nonlinear sensitivity analysis". In: *System modeling and optimization*. Springer, 1982, pp. 762–770.

[146] Krzysztof C Kiwiel. "Convergence and efficiency of subgradient methods for quasiconvex minimization". In: *Mathematical programming* 90.1 (2001), pp. 1–25.

[147] Xiaoyu Li and Francesco Orabona. "On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes". In: vol. 89. Journal of Machine Learning Research (JMLR). PMLR, 2019, pp. 983–992.

[148] Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. "On the almost sure convergence of stochastic gradient descent in non-convex problems". In: *Advances in Neural Information Processing Systems (NIPS)* (2020).

[149] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. *Language models are few-shot learners*. 2020. arXiv: `2005.14165 [cs.CV]`.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[150] Sergey Ioffe and Christian Szegedy. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. 2015. arXiv: `1502.03167 [cs.CV]`.

[151] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting”. In: *Journal of Machine Learning Research (JMLR)* 15.1 (2014), pp. 1929–1958.

[152] Daniel Pierre Bovet, Pierluigi Crescenzi, and D Bovet. *Introduction to the Theory of Complexity*. Prentice Hall London, 1994.

[153] Arthur M Jaffe. “The millennium grand challenge in mathematics”. In: *Notices of the AMS* 53.6 (2006).

[154] Richard M Karp. “Reducibility among combinatorial problems”. In: *Complexity of computer computations*. Springer, 1972.

[155] Hans Mittelmann. *Benchmark of Simplex LP solvers*. 2021. URL: `http://plato.asu.edu/ftp/lpsimp.html`.

[156] Martin Grötschel. *Algorithmische Diskrete Mathematik II–Linear Optimierung*. 2015. URL: `https://www.zib.de/groetschel/teaching/SS2015/Skriptum_ADM_II-2015-07-20.pdf`.

[157] Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1998.

[158] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*. Vol. 6. Athena Scientific Belmont, MA, 1997.

[159] Wikimedia Commons. *P np np-complete np-hard*. [Online, accessed Dezember 2020]. 2007. URL: `https://commons.wikimedia.org/wiki/File:P_np_np-complete_np-hard.svg`.

[160] George B Dantzig. “Maximization of a linear function of variables subject to linear inequalities”. In: *Activity analysis of production and allocation* 13 (1951), pp. 339–347.

[161] Ailsa H. Land and Alison G. Doig. “An automatic method of solving discrete programming problems”. In: *Econometrica: Journal of the Econometric Society* (1960), pp. 497–520.

[162] Ralph E Gomory. “An algorithm for integer solutions to linear programs”. In: *Recent advances in mathematical programming* 64.260-302 (1963), p. 14.

[163] Manfred Padberg and Giovanni Rinaldi. “A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems”. In: *SIAM review* 33.1 (1991), pp. 60–100.

[164] John E. Mitchell. “Branch-and-cut algorithms for combinatorial optimization problems”. In: *Handbook of applied optimization* (2002), pp. 65–77.

[165] LLC Gurobi Optimization. *Gurobi Optimizer Reference Manual*. 2018. URL: `http://www.gurobi.com`.

[166] Martin Grötschel, László Lovász, and Alexander Schrijver. “The ellipsoid method and its consequences in combinatorial optimization”. In: *Combinatorica* 1.2 (1981), pp. 169–197.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[167] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[168] Panos M. Pardalos and Stephen A. Vavasis. “Quadratic programming with one negative eigenvalue is NP-hard”. In: *Journal of Global Optimization* 1.1 (1991), pp. 15–22.

[169] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona. “Fast feature pyramids for object detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 36.8 (2014), pp. 1532–1545.

[170] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2005.

[171] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20.3 (1995), pp. 273–297.

[172] Karen Simonyan and Andrew Zisserman. *Very deep convolutional networks for large-scale image recognition*. 2014. arXiv: `1409.1556 [cs.CV]`.

[173] Fan Yang, Wongun Choi, and Yuanqing Lin. “Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[174] C Lawrence Zitnick and Piotr Dollár. “Edge boxes: Locating object proposals from edges”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2014.

[175] Yoav Freund and Robert E Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[176] Chang Huang, Bo Wu, and Ramakant Nevatia. “Robust object tracking by hierarchical association of detection responses”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2008.

[177] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. “Deepmatching: Hierarchical deformable dense matching”. In: *International Journal of Computer Vision (IJCV)* 120.3 (2016), pp. 300–323.

[178] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. “Joint discriminative and generative learning for person re-identification”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[179] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. *Deep learning for person re-identification: A survey and outlook*. 2020. arXiv: `2001.04193 [cs.CV]`.

[180] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. “Generative adversarial nets”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2014.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[181] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. “Scalable person re-identification: A benchmark”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[182] James Ferryman and Ali Shahrokni. “Pets2009: Dataset and challenge”. In: *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*. 2009.

[183] Ben Benfold and Ian Reid. “Stable multi-target tracking in real-time surveillance video”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

[184] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. “A mobile vision system for robust multi-person tracking”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2008.

[185] Anton Milan, Konrad Schindler, and Stefan Roth. “Challenges of ground truth evaluation of multi-target tracking”. In: *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2013.

[186] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. “Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking”. In: *ECCV Workshop on Benchmarking Multi-Target Tracking (ECCVW)*. Springer. 2016.

[187] Keni Bernardin and Rainer Stiefelhagen. “Evaluating multiple object tracking performance: the CLEAR MOT metrics”. In: *EURASIP Journal on Image and Video Processing* 2008 (2008), pp. 1–10.

[188] Yuan Li, Chang Huang, and Ram Nevatia. “Learning to associate: Hybridboosted multi-target tracker for crowded scene”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009.

[189] Sheng Chen, Alan Fern, and Sinisa Todorovic. “Multi-object tracking via constrained sequential labeling”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[190] Anton Milan, Laura Leal-Taixé, Konrad Schindler, and Ian Reid. “Joint tracking and segmentation of multiple targets”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[191] Thomas Brox and Jitendra Malik. “Object Segmentation by Long Term Analysis of Point Trajectories”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2010.

[192] Roberto Henschel, Laura Leal-Taixé, Bodo Rosenhahn, and Konrad Schindler. *Tracking with multi-level features*. 2016. arXiv: `1607.07304 [cs.CV]`.

[193] Carlo Tomasi and Takeo Kanade. “Detection and tracking of point features”. In: *International Journal of Computer Vision (IJCV)* (1991), pp. 137–154.

[194] Bruce D Lucas, Takeo Kanade, et al. *An iterative image registration technique with an application to stereo vision*. 1981.

[195] Katerina Fragkiadaki and Jianbo Shi. “Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2011.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[196] Katerina Fragkiadaki, Weiyu Zhang, Geng Zhang, and Jianbo Shi. “Two-granularity tracking: mediating trajectory and detections graphs for tracking under occlusions”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2012.

[197] Guillaume Seguin, Piotr Bojanowski, Remi Lajugie, and Ivan Laptev. “Instance-Level Video Segmentation From Object Tracks”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[198] Simon Lacoste-Julien. *Convergence rate of Frank-Wolfe for non-convex objectives.* 2016. arXiv: `1607.00345 [math.OC]`.

[199] Anton Milan, Konrad Schindler, and Stefan Roth. “Multi-target tracking by discrete-continuous energy minimization”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 38.10 (2015), pp. 2054–2068.

[200] Afshin Dehghan and Mubarak Shah. “Binary Quadratic Programing for Online Tracking of Hundreds of People in Extremely Crowded Scenes”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 40.3 (2018), pp. 568–581.

[201] Shayan Modiri Assari, Haroon Idrees, and Mubarak Shah. *Re-identification of Humans in Crowds using Personal, Social and Environmental Constraints*. 2016. arXiv: `1612.02155 [cs.CV]`.

[202] Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. “Block-Coordinate Frank-Wolfe Optimization for Structural SVMs”. In: *ICML*. 2013.

[203] Armand Joulin, Kevin Tang, and Fei-Fei Li. “Efficient image and video co-localization with frank-wolfe algorithm”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer. 2014.

[204] Simon Lacoste-Julien and Martin Jaggi. “On the Global Linear Convergence of Frank-Wolfe Optimization Variants”. In: 2015.

[205] Héctor Allende, Emanuele Frandi, Ricardo Ñanculef, and Claudio Sartori. “Pairwise away steps for the Frank-Wolfe algorithm”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2013.

[206] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[207] Alain Billionnet, Sourour Elloumi, and Amelie Lambert. “Extending the QCR method to general mixed-integer programs”. In: *Mathematical programming* 131.1 (2012), pp. 381–401.

[208] Samuel Burer and Adam N. Letchford. “Non-convex mixed-integer nonlinear programming: A survey”. In: *Surveys in Operations Research and Management Science* 17.2 (2012), pp. 97–106.

[209] Peter L Hammer and Abraham A Rubin. “Some remarks on quadratic programming with 0-1 variables”. In: *Revue française d’informatique et de recherche opérationnelle. Série verte* (1970).

[210] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. “End-to-end people detection in crowded scenes”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[211] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. "Multiple Hypothesis Tracking Revisited: Blending in Modern Appearance Model". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[212] Jiahui Chen, Hao Sheng, Yang Zhang, and Zhang Xiong. "Enhancing Detection Model for Multiple Hypothesis Tracking". In: *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017.

[213] Tino Kutschbach, Erik Bochinski, Volker Eiselein, and Thomas Sikora. "Sequential Sensor Fusion Combining Probability Hypothesis Density and Kernelized Correlation Filters for Multi-Object Tracking in Video Data". In: *IEEE International Conference on Advanced Video and Signal Based Surveillance Workshop (AVSSW)*. 2017.

[214] Xiaoxiao Sun and Liang Zheng. "Dissecting Person Re-identification from the Viewpoint of Viewpoint". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[215] Shaogang Gong, Marco Cristani, Shuicheng Yan, and Chen Change Loy. *Person Re-Identification*. Springer Publishing Company, Incorporated, 2014. ISBN: 1447162951, 9781447162957.

[216] Manon Kok, Jeroen D Hol, and Thomas B. Schön. "Using Inertial Sensors for Position and Orientation Estimation". In: *Foundations and Trends in Signal Processing* 11.1-2 (2017), pp. 1–153.

[217] Antonio R Jimenez, Fernando Seco, Carlos Prieto, and Jorge Guevara. "A comparison of pedestrian dead-reckoning algorithms using a low-cost MEMS IMU". In: *Proceedings of the IEEE International Symposium on Intelligent Signal Processing*. 2009.

[218] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. "Sparse inertial poser: Automatic 3d human pose estimation from sparse imus". In: *Computer Graphics Forum*. Vol. 36. 2. Wiley Online Library. 2017, pp. 349–360.

[219] Stephen Boyd and Lieven Vandenberghe. *Introduction to applied linear algebra: vectors, matrices, and least squares*. Cambridge university press, 2018.

[220] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. "Learning an image-based motion context for multiple people tracking". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[221] Eagle S Jones and Stefano Soatto. "Visual-inertial navigation, mapping and localization: A scalable real-time causal approach". In: *The International Journal of Robotics Research* 30.4 (2011), pp. 407–430.

[222] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. "IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation". In: Georgia Institute of Technology. 2015.

[223] Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. "Human Pose Estimation from Video and IMUs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 38.8 (2016), pp. 1533–1547.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[224] Wenchao Jiang and Zhaozheng Yin. “Combining passive visual cameras and active IMU sensors to track cooperative people”. In: *International Conference on Information Fusion (Fusion).* 2015, pp. 1338–1345.

[225] Wenchao Jiang and Zhaozheng Yin. “Combining passive visual cameras and active IMU sensors for persistent pedestrian tracking”. In: *Journal of Visual Communication and Image Representation* 48 (2017), pp. 419–431.

[226] Thiago Teixeira, Deokwoo Jung, Gershon Dublon, and Andreas Savvides. “Identifying people in camera networks using wearable accelerometers”. In: *Proceedings of the 2nd International Conference on Pervasive Technologies Related to Assistive Environments.* 2009.

[227] Thiago Teixeira, Deokwoo Jung, and Andreas Savvides. “Tasking networked cctv cameras and mobile phones to identify and localize multiple people”. In: *Proceedings of the 12th ACM international conference on Ubiquitous computing.* 2010.

[228] Takashi Hamatani, Yudai Sakaguchi, Akira Uchiyama, and Teruo Higashino. “Player identification by motion features in sport videos using wearable sensors”. In: *International Conference on Mobile Computing and Ubiquitous Networking (ICMU).* IEEE. 2016.

[229] Massimo Camplani, Adeline Paiement, Majid Mirmehdi, Dima Damen, Sion Hannuna, Tilo Burghardt, and Lili Tao. “Multiple human tracking in RGB-depth data: a survey”. In: *IET computer vision* 11.4 (2016), pp. 265–285.

[230] Alexandre Alahi, Albert Haque, and Li Fei-Fei. “RGB-W: When vision meets wireless”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2015.

[231] Peter Nillius, Josephine Sullivan, and Stefan Carlsson. “Multi-Target Tracking–Linking Identities using Bayesian Network Inference”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2006.

[232] Kenneth Levenberg. “A method for the solution of certain non-linear problems in least squares”. In: *Quarterly of applied mathematics* 2.2 (1944), pp. 164–168.

[233] Donald W Marquardt. “An algorithm for least-squares estimation of nonlinear parameters”. In: *Journal of the society for Industrial and Applied Mathematics* 11.2 (1963), pp. 431–441.

[234] *XSens.* `https://www.xsens.com/products/`. [Online; accessed June, 2016]. (Visited on 10/12/2016).

[235] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. “Microsoft coco: Common objects in context”. In: *Proceedings of the European Conference on Computer Vision (ECCV).* Springer. 2014.

[236] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. “Part-aligned bilinear representations for person re-identification”. In: *Proceedings of the European Conference on Computer Vision (ECCV).* Springer. 2018.

[237] Tijmen Tieleman and Geoffrey Hinton. *Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude.* COURSERA: Neural Networks for Machine Learning. 2012.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[238] Péter Kovács. “Minimum-cost flow algorithms: an experimental evaluation”. In: *Optimization Methods and Software* 30.1 (2015), pp. 94–127.

[239] JW Suurballe. “Disjoint paths in a network”. In: *Networks* 4.2 (1974), pp. 125–145.

[240] Tali Eilam-Tzoreff. “The disjoint shortest paths problem”. In: *Discrete Applied Mathematics* 85.2 (1998), pp. 113–138.

[241] Shimon Even, Alon Itai, and Adi Shamir. “On the Complexity of Timetable and Multicommodity Flow Problems”. In: *SIAM J. Comput.* 5 (Dec. 1976), pp. 691–703.

[242] Torsten Tholey. “Linear time algorithms for two disjoint paths problems on directed acyclic graphs”. In: *Theoretical Computer Science* 465 (2012), pp. 35–48.

[243] Sebastian. Nowozin and Christoph H. Lampert. “Global Interactions in Random Field Models: A Potential Function Ensuring Connectedness”. In: *SIAM Journal on Imaging Sciences* 3.4 (2010), pp. 1048–1074.

[244] Margret Keuper, Evgeny Levinkov, Nicolas Bonneel, Guillaume Lavoué, Thomas Brox, and Bjoern Andres. “Efficient Decomposition of Image and Mesh Graphs by Lifted Multicuts”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[245] Andrea Horňáková, Jan-Hendrik Lange, and Bjoern Andres. “Analysis and Optimization of Graph Decompositions by Lifted Multicuts”. In: *International Conference on Machine Learning (ICML)*. 2017.

[246] Thorsten Beier, Constantin Pape, Nasim Rahaman, Timo Prange, Stuart Berg, Davi D Bock, Albert Cardona, Graham W Knott, Stephen M Plaza, Louis K Scheffer, et al. “Multicut brings automated neurite segmentation closer to human performance”. In: *Nature Methods* 14.2 (2017), p. 101.

[247] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. “Instancecut: from edges to instances with multicut”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5008–5017.

[248] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn. “Branch-and-price global optimization for multi-view multi-target tracking”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2012.

[249] Margret Keuper, Siyu Tang, Yu Zhongjie, Bjoern Andres, Thomas Brox, and Bernt Schiele. “A multi-cut formulation for joint segmentation and tracking of multiple objects”. In: *arXiv preprint arXiv:1607.06317* (2016).

[250] Ratnesh Kumar, Guillaume Charpiat, and Monique Thonnat. “Multiple object tracking by efficient graph partitioning”. In: *Asian Conference on Computer Vision*. Springer. 2014, pp. 445–460.

[251] Ergys Ristani and Carlo Tomasi. “Tracking multiple people online and in real time”. In: *Asian Conference on Computer Vision*. Springer. 2014, pp. 444–459.

[252] Maryam Babaee, Ali Athar, and Gerhard Rigoll. “Multiple People Tracking Using Hierarchical Deep Tracklet Re-identification”. In: *arXiv preprint arXiv:1811.04091* (2018).

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[253] Weiming Hu, Xinchu Shi, Zongwei Zhou, Junliang Xing, Haibin Ling, and Stephen Maybank. “Dual L1-Normalized Context Aware Tensor Power Iteration and Its Applications to Multi-object Tracking and Multi-graph Matching”. In: *International Journal of Computer Vision (IJCV)* (Oct. 2019). ISSN: 1573-1405. URL: `https://doi.org/10.1007/s11263-019-01231-y`.

[254] Stephen A Cook. “The complexity of theorem-proving procedures”. In: *Proceedings of the third annual ACM symposium on Theory of computing*. ACM. 1971, pp. 151–158.

[255] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. “Person transfer gan to bridge domain gap for person re-identification”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 79–88.

[256] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. “DeepFlow: Large displacement optical flow with deep matching”. In: *IEEE Intenational Conference on Computer Vision*. Sydney, Australia, Dec. 2013. URL: `http://hal.inria.fr/hal-00873592`.

[257] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. “Exploit the connectivity: Multi-object tracking with trackletnet”. In: *ACM International Conference on Multimedia*. 2019, pp. 482–490.

[258] Hao Sheng, Yang Zhang, Jiahui Chen, Zhang Xiong, and Jun Zhang. “Heterogeneous association graph fusion for target association in multiple object tracking”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.11 (2018), pp. 3269–3280.

[259] Long Chen, Haizhou Ai, Rui Chen, and Zijie Zhuang. “Aggregate Tracklet Appearance Features for Multi-Object Tracking”. In: *IEEE Signal Processing Letters* 26.11 (2019), pp. 1613–1617.

[260] Liqian Ma, Siyu Tang, Michael J Black, and Luc Van Gool. “Customized multi-person tracker”. In: *Asian Conference on Computer Vision*. Springer. 2018, pp. 612–628.

[261] Peng Chu, Heng Fan, Chiu C Tan, and Haibin Ling. “Online multi-object tracking with instance-aware tracker and dynamic model refreshment”. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2019.

[262] Long Chen, Haizhou Ai, Chong Shang, Zijie Zhuang, and Bo Bai. “Online multi-object tracking with convolutional neural networks”. In: *Proceedings of the IEEE Conference on Image Processing (ICIP)*. 2017, pp. 645–649.

[263] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. “The Secrets of Salient Object Segmentation”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2014.

[264] Jun Xiang, Guohan Xu, Chao Ma, and Jianhua Hou. “End-to-end learning deep crf models for multi-object tracking”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[265] Patrick Schramowski, Christian Bauckhage, and Kristian Kersting. “Neural conditional gradients”. In: *arXiv preprint arXiv:1803.04300* (2018).

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

[266] Marin Vlastelica Pogančić, Anselm Paulus, Vit Musil, Georg Martius, and Michal Rolinek. “Differentiation of blackbox combinatorial solvers”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[267] Xi Sun, Xinshuo Weng, and Kris Kitani. “When We First Met: Visual-Inertial Person Localization for Co-Robot Rendezvous”. In: *IROS* (2020).

[268] Paul Swoboda and Bjoern Andres. “A message passing algorithm for the minimum cost multicut problem”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[269] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. “Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[270] Shu Wang, Huchuan Lu, Fan Yang, and Ming-Hsuan Yang. “Superpixel tracking”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE. 2011.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

VDI verlag

Alle 23 Reihen der „Fortschritt-Berichte VDI“
in der Übersicht – bequem recherchieren unter:
elibrary.vdi-verlag.de

Und direkt bestellen unter:
www.vdi-nachrichten.com/shop

Reihe 01	Konstruktionstechnik/ Maschinenelemente
Reihe 02	Fertigungstechnik
Reihe 03	Verfahrenstechnik
Reihe 04	Bauingenieurwesen
Reihe 05	Grund- und Werkstoffe/Kunststoffe
Reihe 06	Energietechnik
Reihe 07	Strömungstechnik
Reihe 08	Mess-, Steuerungs- und Regelungstechnik
Reihe 09	Elektronik/Mikro- und Nanotechnik
Reihe 10	Informatik/Kommunikation
Reihe 11	Schwingungstechnik
Reihe 12	Verkehrstechnik/Fahrzeugtechnik
Reihe 13	Fördertechnik/Logistik
Reihe 14	Landtechnik/Lebensmitteltechnik
Reihe 15	Umwelttechnik
Reihe 16	Technik und Wirtschaft
Reihe 17	Biotechnik/Medizintechnik
Reihe 18	Mechanik/Bruchmechanik
Reihe 19	Wärmetechnik/Kältetechnik
Reihe 20	Rechnergestützte Verfahren
Reihe 21	Elektrotechnik
Reihe 22	Mensch-Maschine-Systeme
Reihe 23	Technische Gebäudeausrüstung

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

OHNE PROTOTYP GEHT NICHTS IN SERIE.

Unser Podcast ist das Werkzeug, mit dem Sie Ihre Karriere in allen Phasen entwickeln – vom Studium bis zum Chefsessel. Egal, ob Sie Ingenieur*in, Mechatroniker*in oder Wissenschaftler*in sind: Prototyp begleitet Sie. Alle 14 Tage hören Sie die Redaktion von INGENIEUR.de und VDI nachrichten im Gespräch mit prominenten Gästen.

INGENIEUR.de
TECHNIK - KARRIERE - NEWS

PROTOTYP
Karriere-Podcast

JETZT REINHÖREN UND KOSTENFREI ABONNIEREN:
WWW.INGENIEUR.DE/PODCAST

IN KOOPERATION MIT VDI NACHRICHTEN

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.

REIHE 10
INFORMATIK/
KOMMUNIKATION

NR. 875

ISBN 987-3-18-387510-7

BAND
1|1

VOLUME
1|1

https://doi.org/10.51202/9783186875105 - Generiert durch IP 216.73.216.60, am 24.01.2026, 03:56:38. © Urheberrechtlich geschützter Inhalt. Ohne gesonderte
Erlaubnis ist jede urheberrechtliche Nutzung untersagt, insbesondere die Nutzung des Inhalts im Zusammenhang mit, für oder in KI-Systemen, KI-Modellen oder Generativen Sprachmodellen.