Accurate Long-Term Multiple People Tracking Using Video and Body-Worn IMUs

Roberto Henschel, Timo von Marcard, and Bodo Rosenhahn

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 inertial measurement units (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 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. Since no dataset exists for this new setting, 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.**

Index Terms— **Multiple people tracking, graph labeling, sensor fusion, IMU, human motion analysis.**

I. Introduction

MULTIPLE people tracking (MPT) in image sequences has been an active field of research for decades. Several applications exist where trajectories are required for further analysis and interpretation. This could be to understand social interactions of humans [1]–[4], support urban planning [5], secure areas against dangerous behavior [6] or to provide an automatic analysis of a player's performance in sports [7]–[10]. Most state-of-the-art MPT approaches tackle this problem in two steps: First, a person detector is applied to each frame of the image sequence. Then, an optimization problem is solved, which clusters all detections such that ideally each cluster represents the trajectory of a person, and false detections remain unconsidered.

A crucial part of this strategy is to derive a measure of whether two detections belong to the same person or not. Typically, this involves a motion or appearance model. A motion model attempts to assign likelihoods to observed person movements. This is very generic and only depends on the corner coordinates of the detection boxes. However, as soon as the motion becomes more dynamic, simple motion models [11] are insufficient and the tracking accuracy degrades. In particular, most motion models assume low and constant velocities, which holds for pedestrians only within a short temporal window [12]. Another complementary strategy is to model relations between detections based on the 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 this information exclusively [12], [13] or in combination with motion models [14], [15]. A major advantage of utilizing appearance information over motion models is that they allow to relate detections that are temporally far apart. This facilitates 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, where team members wear almost identical jerseys. Further challenges arise in cases of low-resolution images, changes in the viewpoint [16] or lighting conditions [17], or if people change appearance throughout a sequence, e.g. they put on a jacket or open an umbrella. Then, the assumption of appearance constancy is violated, and the tracking accuracy degrades consequently.

In this work, we propose to complement visual information from video with motion information from body-worn inertial measurement units (IMUs). IMUs are small motion sensors measuring 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 or not available. The latter could be the case if people wear team jerseys or uniforms, if night-vision

Manuscript received October 30, 2019; revised June 5, 2020; accepted July 17, 2020. Date of publication August 13, 2020; date of current version August 25, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianbing Shen. *(Corresponding author: Roberto Henschel.)*

The authors are with the TNT Group, Leibniz University of Hannover, 30167 Hannover, Germany (e-mail: henschel@tnt.uni-hannover.de).

Digital Object Identifier 10.1109/TIP.2020.3013801

is used or if processing person appearance is prohibited due to privacy regulations. Further, 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.

Incorporating additional sensory input for the task of MPT creates a very different problem setup compared to the aforementioned vision-only methods. In particular, this involves (i) solving the data association problem of detections to trajectories in the 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. On the other hand, this fusion allows to combine the strength of two complementary input sources, and is thus a promising concept to obtain highly accurate trajectories. We denote this new 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 [18], [19]. If this were 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 the 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) [20], to image data being a two-dimensional pixel array. Further, IMU measurements often fit to several people at a time step, and the person wearing the IMU might be occluded or out of the camera view.

A. Contributions

Vision-based multiple people tracking systems rely on certain assumptions about the motion and the appearance of the objects to be tracked. Once these assumptions are violated, the tracking accuracy degrades. This frequently happens if people wear similar apparel, or if persons are tracked across different recordings, e.g. in a long-term motion study. 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 consider a setting where people are wearing IMU sensors, allowing to tackle tracking and identification holistically. In particular, this setting allows one (i) to reduce the dependency on artificial motion models (velocities in the 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. To this end, we propose a graph labeling formulation that fuses the video signal with complementary motion information from the IMU sensors.

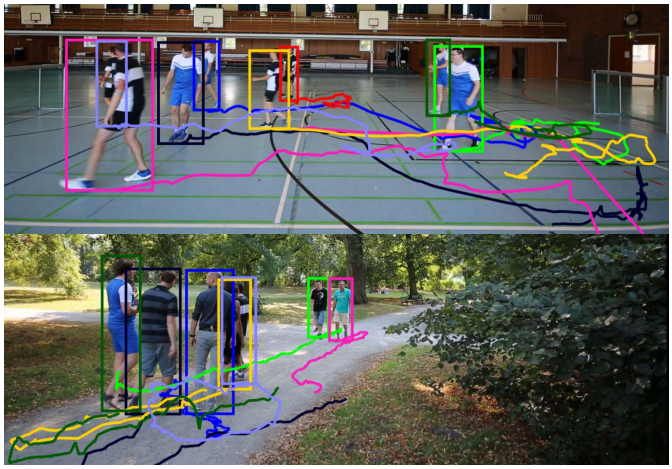

Fig. 1. Qualitative results obtained by fusing person detections and local motion measurements of body-worn IMUs. Instead of relying on appearance information, the proposed approach enables accurate long-term tracking by finding a globally optimal assignment of detection boxes to IMU devices, such that resultant trajectories in the video are consistent with the IMU measurements.

The underlying idea is that a global assignment of detection boxes in the video to specific labels representing person identities has to 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. Finally, a globally consistent assignment is obtained by solving a binary quadratic problem.

Unfortunately, no existing dataset for multiple-people tracking contains body-worn IMU data. To evaluate our proposed tracking approach, we created a new dataset. A special emphasis was put on similar person appearance, heavy occlusions, and non-linear motion. These are the situations in which model assumptions implicitly used in vision-based approaches are violated. Also, since such tracking scenarios are currently missing in standard benchmarks such as [21], [22], the new dataset could be valuable to evaluate and improve vision-only approaches.

The present work is an extension of our preliminary conference paper [23] and improves it in several ways:

- We introduce and evaluate an interpolation method in Section III-E that uses IMU data to recover accurate people trajectories if visual information is missing due to occlusions. This demonstrates a unique advantage of incorporating IMUs to the task of multiple people tracking.
- We extend the evaluation of our tracking approach in Section IV-D. Two new trackers are presented. One uses

orientation only, while the other additionally uses the acceleration signal.
- We provide an extensive evaluation of the orientation predictor and validate the necessity of the perspective correction in Section IV-D.
- We propose a new dataset, *VIMPT2019*, that contains video and body-worn IMU data for challenging soccer and outdoor scenes. The dataset contains detection boxes and ground-truth labels, and is publicly available for research purposes.[1]

Overall, the entire work contains the following contributions:

- We introduce a new extension to the multiple people tracking problem, termed Video Inertial Multiple People Tracking (VIMPT), which augments the video recording with inertial measurements from body-worn IMUs.
- We release the first dataset for the VIMPT setting.
- We present the first multiple people tracking system, which is able to fuse video information with IMU measurement for the purpose of VIMPT.
- We design a simple, yet effective neural network, which relates detections to orientation measurements by utilizing a novel perspective correction.

II. Related Work

A. Data Association

Most multiple people tracking works employ the tracking-by-detection paradigm [13]–[15], [24]–[32] that connects either detections [13], [15], [28], [32] or precomputed tracklets [26], [29], [33], [34] to form trajectories. The problem of creating trajectories, often denoted as data association, is usually formulated as a graph optimization problem. Several works apply network-flow [11], [35], while more recently submodular optimization [36], minimum cost multicut [12], [15], [27], lifted disjoint paths [32] or graph labeling [13] formulations have been proposed.

B. Association Weights

Crucial for graph-based tracking approaches are the association weights between detections (or tracklets) that indicate how likely they belong to the same person. Several works have focused on obtaining these weights from motion models [11], [33], [37]–[40]. Typically, a linear constant velocity model within short time windows is assumed [11], [38]. However, the performance of these approaches degrades if motions become more dynamic or people get temporarily occluded. Consequently, current state-of-the-art tracking systems [12]–[15], [27], [41]–[45] rely on appearance models that are invariant to these issues. They use sophisticated neural networks to derive association weights from the visual information. Some works derive attention weights [45] that indicate whether the appearance information is reliable. The accuracy of appearance features has improved to a level that some works reformulate the tracking problem as a person re-identification problem [14], [44]. Accordingly, some works compose a multiple people tracking system out of multiple visual object trackers [44]. Visual object trackers [46]–[48] require an initial mask of the object to be tracked, and essentially detect the object in each frame. During this process, the appearance model of the object is updated frequently.

Despite the impressive progress of current tracking methods that build upon appearance models, common to all these 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.

An alternative solution is to integrate additional modalities into the tracking method.

C. Vision and Inertial Sensors

Body-worn inertial sensors provide motion information independent of the visibility of the persons. However, it is not possible to recover the 3D person trajectory from IMU information alone [18], [19]. In contrast, using the video signal allows to extract positional information, which is complementary to the IMU motion information.

Consequently, IMUs have been combined with visual information in many applications, e.g. fusing video and inertial data to stabilize self localization and mapping (SLAM) [49], [50]. The same modalities have been used to recover human poses [51], [52].

There exist only very few works that incorporate IMUs for people tracking in videos [53]–[55]. The closest reference to our work is [54], which tackles single person tracking. In this work, an IMU-equipped person has to be manually localized in the first video frame. Then, IMU information are used to recover the trajectory in situations where the visual tracker fails. Instead, we propose a method that automatically identifies and tracks multiple IMU-equipped persons. In addition, our fusion formulation uses both modalities simultaneously in the optimization problem of the tracker, thereby combining the advantages of both sensors.

D. Other Sensor Modalities

While we are the first that combine video information with inertial sensors for the purpose of multiple people tracking, there exist several works that incorporate other sensor modalities, e.g. Camplani *et al.* [56] provide a survey of tracking approaches using RGB-D cameras. However, depth cameras work only indoors and have a limited depth range. Another work [57] 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 VIMPT setting does not suffer from these limitations.

E. 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 [36], [58],

[1] Access to the dataset via http://www.tnt.uni-hannover.de/project/VIMPT/

[59], for understanding social behavior of humans [2]–[4], or in order to assess the performance of athletes [7]–[10]. 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 throughout a recording but also across different recordings.

Manual labeling of trajectories is a tedious task. Hence, several works are focused on automatically obtaining the true identities but consider this as a post-processing step. For instance, the label computation can be formulated as a Bayesian network inference problem [60], where the computed trajectories belonging to the same person are equally labeled by defining some measure of how likely two trajectories belong to the same person. This 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 the recording, both tasks can be tackled [8]. Then, trajectories are labeled by employing a conditional random field using the visual features from the database. In contrast to these visual approaches, the VIMPT setting allows to simultaneously seek for a solution that is consistent with the video and IMU labels, so that labeling all detections is not only a desired task but also helps to obtain accurate tracking results.

III. Method

We follow the tracking-by-detection paradigm and group detections to short tracklets in a first step. The tracking task can then be formulated to assign IDs to tracklets, such that all tracklets with identical IDs correspond to person trajectories in the video.

In the context of this work, we solve the tracking task (or data association) by incorporating motion information from body-worn IMUs. Hence, we formulate a graph labeling problem 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.

We integrate the IMU signals at different conceptual levels: For each potential detection to IMU assignment, we require that the person orientation, as seen by the camera, is consistent with the corresponding IMU orientation. Orientation consistency alone is very ambiguous, and 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. Figure 2 illustrates the graph and shows an exemplary labeling solution. We refer the interested reader to [61] for more information on IMUs and corresponding orientation and acceleration signals.

A. Model

The tracking task is formulated using an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{C}, \mathcal{L})$, where $\mathcal{V}$ is the vertex set comprising 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 $l \in \mathcal{L}$, where the label

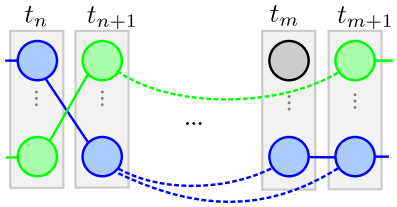

Fig. 2. Every tracklet represents a node in the graph. Each node can be assigned to an IMU device (indicated by color) and is linked to other nodes by short-term edges (solid) and long-term edges (dashed). An edge is activated if corresponding nodes share the same color. The idea is that every graph 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.

set $\mathcal{L} = \{1, 2, 3, \ldots, P\}$ contains an ID for all P persons wearing an IMU. We represent each detection $\mathbf{d}$ as a 5-dimensional vector $(x, y, w, h, t_\mathbf{d}) \in \mathbb{R}^5$, where (x, y) denotes the position at the middle of the lower edge of the detection box in image coordinates (also called *foot position*), w and h represent the box width and height, respectively, and $t_\mathbf{d}$ holds the timestamp of detection $\mathbf{d}$. Each tracklet υ represents a set of detections that do not overlap in time and are spatio-temporally consistent.

At this point, we introduce the notion of an assignment hypothesis $\mathcal{H} = (\upsilon, l)$, which associates a label $l \in \mathcal{L}$ to tracklet $\upsilon \in \mathcal{V}$. Associated to each hypothesis are assignment costs $c_\upsilon^l \in \mathcal{C}$ reflecting the assignment likelihood and indicator variables x_υ^l, which take value 1 if hypothesis $\mathcal{H}$ is selected, and 0 otherwise. Additionally, for pairs of hypotheses sharing the same label and whose vertices are connected by an edge $e \in \mathcal{E}$, we consider compatibility costs $c_e^l \in \mathcal{C}$ modeling the likelihood that two tracklets belong to the same person.

The tracking task is then to select hypotheses for the entire sequence that minimize the total costs. This can be cast into a binary optimization problem:

$$\underset{\mathbf{X} \in \mathcal{F} \cap \{0,1\}^{|\mathcal{V}| \times P}}{\arg\min} \sum_{l \in \{1,\ldots,P\}} \left(\sum_{\upsilon \in \mathcal{V}} c_\upsilon^l x_\upsilon^l + \sum_{e \in \mathcal{E}} c_e^l \prod_{\upsilon \in e} x_\upsilon^l \right), \quad (1)$$

where the entry of matrix $\mathbf{X}$ at row l and column corresponding to υ equals x_υ^l. The feasibility set $\mathcal{F}$ is subject to

$$\forall \upsilon \in \mathcal{V} : \sum_{l=1}^{P} x_\upsilon^l \leq 1, \quad (2)$$

$$\forall t \in \{1, \ldots, T\} \, \forall l \in \{1, \ldots, P\} : \sum_{\upsilon \in \mathcal{V}_t} x_\upsilon^l \leq 1. \quad (3)$$

The subset $\mathcal{V}_t \subset \mathcal{V}$ comprises all tracklets υ that contain a detection in frame t and T denotes the number of image frames. Eq. (2) ensures that each tracklet υ is assigned to at most one label and Eq. (3) guarantees that a label is not assigned to more than one tracklet at a time.

The solution of (1) provides the desired trajectories. Specifically, for each label $l \in \mathcal{L}$, we obtain a trajectory $\mathcal{T}_l : T(l) \to \mathbb{R}^5$, where $T(l) := \{t_\mathbf{d} \mid \mathbf{d} \in \upsilon, x_\upsilon^l = 1\}$ is the set of frames at which a detection has been assigned to person l, and $\mathcal{T}_l(t) := \mathbf{d}$ is the assigned detection $\mathbf{d}$ at time t.

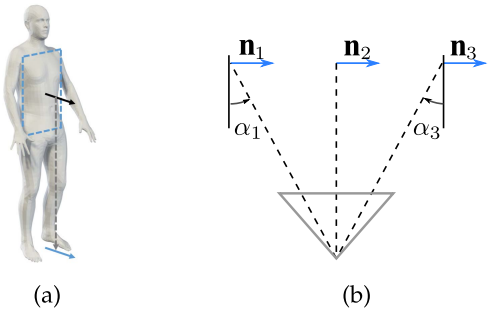

Fig. 3. (a) We define person orientation in terms of the normal vector of the torso's coronal plane (black arrow) projected to the ground plane (blue arrow). (b) Consider a top view of a person walking on a straight line parallel to the image plane. Even though the person's torso orientation $\mathbf{n}$ 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.

Next, we describe the unary and pairwise potentials in detail. Specifically, we introduce consistency features which are mapped to costs c_v^l and c_e^l, as described in Section IV-C.

B. Unary Features

In order to provide a measurement for the likelihood of an assignment hypothesis $\mathcal{H} = (v, l)$, we estimate the orientations of a person in each detection box of tracklet v and compare those orientations to the temporally aligned orientation measurements of IMU l.

We define the person orientation $\mathbf{n} \in \mathbb{R}^2$ as the normal vector of the torso's coronal plane projected to the ground plane, as illustrated in Figure 3(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 detection $\mathbf{d}$, we seek to estimate the heading $\hat{\mathbf{n}}_\mathbf{d}$ of the person. However, the observed heading in $I_\mathbf{d}$ depends on the person position in the image, see Figure 3(b). 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. We compensate for this by considering 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 box position $\mathbf{p}_\mathbf{d}$, and the depth-axis of the camera. In order to compensate the perspective influences, we rotate the perceived orientation by $-\alpha_\mathbf{d}$ and obtain the prediction $\hat{\mathbf{n}}_\mathbf{d}$, compare Figure 3(b).

In order to obtain the person heading from image data, we employ a neural net to learn the mapping $I_\mathbf{d} \mapsto \hat{\mathbf{n}}_\mathbf{d}$. More specifically, we extend VGG16 [62] pretrained on ImageNet [63] to regress the heading, which also incorporates the aforementioned perspective correction (PC) in the last layer. We refer to this network as the Visual Heading Network

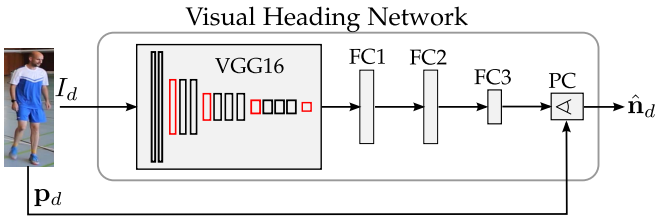

Fig. 4. The Visual Heading Network predicts the heading $\hat{\mathbf{n}}_\mathbf{d}$ of a person using the image data $I_\mathbf{d}$ of detection $\mathbf{d}$. Based on the box position $\mathbf{p}_\mathbf{d}$, the network performs a perspective correction (PC) in the last layer.

(VHN) in the following. A graphical illustration showing the network architecture is depicted in Figure 4. In the VIMPT setting, IMUs are consistently placed on the back of each person such that the local sensor z-axis corresponds to the normal vector of the torso's coronal plane. Hence, we get the measured torso orientation vector $\mathbf{n}_{l,t}$ of IMU l at time t according to

$$\mathbf{n}_{l,t} = \Pi(\mathbf{R}_{l,t}\mathbf{z}), \tag{4}$$

where $\mathbf{z} = [\,0\,0\,1\,]^T$ is the local z-axis vector, $\mathbf{R}_{l,t} \in SO(3)$ is the measured IMU orientation mapping the local sensor coordinate frame to the global coordinate frame, and $\Pi : \mathbb{R}^3 \to \mathbb{R}^2$ 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 hypothesis $\mathcal{H}$ as

$$f_{\text{ori}}(\mathcal{H}) = \frac{1}{N_v} \sum_{\mathbf{d} \in v} \Phi(\hat{\mathbf{n}}_\mathbf{d}, \mathbf{n}_{l,t_\mathbf{d}}), \tag{5}$$

where Φ denotes the cosine similarity

$$\Phi(\mathbf{n}_1, \mathbf{n}_2) := \frac{\mathbf{n}_1 \cdot \mathbf{n}_2}{\|\mathbf{n}_1\| \, \|\mathbf{n}_2\|} \in [-1, 1] \tag{6}$$

between vectors $\mathbf{n}_1, \mathbf{n}_2 \in \mathbb{R}^2$, N_v corresponds to the number of detections of tracklet v, and $t_\mathbf{d}$ represents the time stamp of a detection $\mathbf{d}$. The orientation feature $f_{\text{ori}}(\mathcal{H})$ thus measures the average orientation consistency of the tracklet v to the IMU device with ID l.

C. Pairwise Features

We define pairwise features, which represent the compatibility of two hypotheses $\mathcal{H} = (v, l)$ and $\mathcal{H}' = (v', l)$. Two hypotheses are said to be compatible, if the assignment of a joint label l to v and v' is reasonable with respect to spatio-temporal aspects.

1) Spatio-Temporal Features: Within a short temporal window, a person cannot move arbitrarily fast. Hence, the tracklets of a compatible hypothesis pair should be spatially close, and corresponding detection boxes should be similar in size. We derive corresponding features in the following.

For each detection box $\mathbf{d}$, we obtain an estimate $\prod_{\text{Foot}}(\mathbf{d}) := \mathbf{p}_\mathbf{d} \in \mathbb{R}^2$ of the foot position of the corresponding person in world coordinates (on the ground plane) by projecting the foot position of the detection to the ground plane of the

scene. Hence, for detections $\mathbf{d}$ of v and $\mathbf{d}'$ of v', let $\mathbf{v}_{3D}(\mathbf{d}, \mathbf{d}')$ denote the velocity in 3D from $\mathbf{d}$ to $\mathbf{d}'$. Let $N(v, v')$ be the set of all pairs of detections between $\mathcal{H}$ and $\mathcal{H}'$. We define the mean velocity feature between $\mathcal{H}$ and $\mathcal{H}'$ as

$$f_{\text{vel}}(\mathcal{H}, \mathcal{H}') = \frac{1}{|N(v, v')|} \sum_{(\mathbf{d},\mathbf{d}') \in N(v,v')} \left\| \mathbf{v}_{3D}(\mathbf{d}, \mathbf{d}') \right\|_2. \quad (7)$$

Additionally, we compare the detection box heights included in both hypotheses. Let $h_{\mathbf{d}}$ denote the height of detection box $\mathbf{d}$ in pixels. We define a compatibility measure $f_h(\mathbf{d}, \mathbf{d}')$ based on the heights of detections $\mathbf{d}$ and $\mathbf{d}'$ according to

$$f_h(\mathbf{d}, \mathbf{d}') = f_t(\mathbf{d}, \mathbf{d}') \frac{|h_{\mathbf{d}} - h_{\mathbf{d}'}|}{\min\{h_{\mathbf{d}}, h_{\mathbf{d}'}\}}, \quad (8)$$

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 k frames apart, we set

$$f_t(\mathbf{d}, \mathbf{d}') = \frac{1}{\log(c + k)}, \quad (9)$$

where c is chosen such that $\log(c+1) = 1$. Then, $f_t(\mathbf{d}, \mathbf{d}') = 1$ holds, if the frame distance between $\mathbf{d}$ and $\mathbf{d}'$ is 1 and it increases slowly with bigger frame distances k. Finally, we define our box height feature as

$$f_{\text{height}}(\mathcal{H}, \mathcal{H}') = \frac{1}{|N(v, v')|} \sum_{(\mathbf{d},\mathbf{d}') \in N(v,v')} f_h(\mathbf{d}, \mathbf{d}'). \quad (10)$$

Both f_{vel} and f_{height} are features which are meaningful within short temporal windows. However, in this work, we focus on sequences where people get occluded or fall out of the camera view quiet often and for longer time periods. Hence, in the following, we utilize acceleration measurements to link hypotheses that cover larger temporal horizons.

2) Acceleration Feature: Ideally, the ground position $\mathbf{p}_{t_1} \in \mathbb{R}^2$ at time t_1 of an IMU can be recovered by double integration of the corresponding acceleration signal $\mathbf{a}$ according to

$$\mathbf{p}_{t_1} = \mathbf{p}_{t_0} + \mathbf{v}_{t_0}(t_1 - t_0) + \int_{t_0}^{t_1} \int_{t_0}^{u} \mathbf{a}(s) ds\, du, \quad (11)$$

where t_0, $\mathbf{p}_{t_0}$, and $\mathbf{v}_{t_0}$ denote initial time, initial position, and initial velocity, respectively. Please note that $\mathbf{a}$ in this case represents the gravity-free acceleration in global coordinates.

Now let $\mathbf{p}_{t_0}$ denote the 2D ground position of detection $\mathbf{d}$ and $\mathbf{p}_{t_1}$ the 2D ground position of $\mathbf{d}'$. After double integration of the acceleration signal, we can solve Eq. (11) for the initial velocity, which we denote $\mathbf{v}_{\text{IMU}}(\mathbf{d}, \mathbf{d}')$. Thus,

$$\mathbf{v}_{\text{IMU}}(\mathbf{d}, \mathbf{d}') = \frac{\mathbf{p}_{t_1} - \mathbf{p}_{t_0} - \int_{t_0}^{t_1} \int_{t_0}^{u} \mathbf{a}(s) ds\, du}{(t_1 - t_0)}. \quad (12)$$

Concurrently, we can approximate the velocity $\mathbf{v}_{\mathbf{d}}$ of a person at initial time t_0 in terms of finite differences of neighboring detections of $\mathbf{d}$. Hence, for a compatible hypotheses pair $\mathcal{H}$ and $\mathcal{H}'$, the velocity differences

$$f_v(\mathbf{d}, \mathbf{d}') = \left\| \mathbf{v}_{\text{IMU}}(\mathbf{d}, \mathbf{d}') - \mathbf{v}_{\mathbf{d}} \right\|_2 \quad (13)$$

should be small for all possible detection pairs $\mathbf{d} \in v$ and $\mathbf{d}' \in v'$. We define the acceleration feature as the set of all such differences according to

$$f_{\text{acc}}(\mathcal{H}, \mathcal{H}') = \left\{ f_v(\mathbf{d}, \mathbf{d}') \mid (\mathbf{d}, \mathbf{d}') \in N(v, v') \right\}. \quad (14)$$

D. Optimization

The graph labeling problem defined in (1) is a binary quadratic program. We reformulate this program as an equivalent binary linear program (BLP) by introducing slack variables: Each product of variables $x_v^l x_{v'}^l$ is replaced by a new variable $z_{v,v'}^l$, and the following constraints are added:

$$z_{v,v'}^l \leq x_v^l, \quad (15)$$

$$z_{v,v'}^l \leq x_{v'}^l, \quad (16)$$

$$z_{v,v'}^l \geq x_v^l + x_{v'}^l - 1, \quad (17)$$

$$z_{v,v'}^l \in \{0, 1\}. \quad (18)$$

A similar reformulation is proposed in [64]. The resulting problem can then be solved to optimality using BLP solvers like gurobi [65].

E. Interpolation

The solution to problem (1) assigns detection boxes to unique IMU devices. Given these associations, it becomes feasible to accurately reconstruct the trajectory of a person using IMU accelerations. In particular, the positions of a person can be recovered in all frames, even though 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 $\mathcal{T}$ tracking a person with label l and define $\text{proj}(\mathcal{T}) := \prod_{\text{Foot}} \circ \mathcal{T}$ to be the foot coordinates of the trajectory projected to the ground plane, described in world coordinates. Further, let t_{F} and t_{L} denote the timestamps of the first and last detection assigned to trajectory $\mathcal{T}$, respectively. Then, the interpolation has to recover the locations for all image frames in $[t_{\text{F}}, t_{\text{L}}]$.

Given the correspondences between visual information and IMU device l, the obtained trajectory $\mathcal{T}$ is improved by our interpolation algorithm by seeking for the interpolated trajectory $\mathcal{T}_{\text{int}}$ that is spatially close to $\mathcal{T}$ while following the motion information given by IMU sensor l (in terms of acceleration).

Let $\mathfrak{T}$ denote the set of all possible trajectories in this time window. Then, we seek for a trajectory that minimizes the following optimization problem:

$$\min_{\mathcal{T}' \in \mathfrak{T}} \Big(w\, e_{\text{pos}}(\mathcal{T}', \text{proj}(\mathcal{T}), l) + (1 - w)\, e_{\text{acc}}(\mathcal{T}', \mathbf{a}_l, l) \Big), \quad (19)$$

where the residual

$$e_{\text{pos}}(\mathbf{p}, \mathbf{q}, l) = \sum_{t \in T(l)} \frac{\|\mathbf{p}(t) - \mathbf{q}(t)\|_2^2}{|T(l)|} \quad (20)$$

measures the mean squared distance between the interpolated trajectory and the detections of trajectory $\mathcal{T}$. The residual

$$e_{\text{acc}}(\mathbf{p}, \mathbf{a}, l) = \sum_{t \in T_{\text{IMU}}(l)} \frac{\left\| \mathbf{a}(t) - \hat{\mathbf{a}}(\mathbf{p}, t) \right\|_2^2}{|T_{\text{IMU}}(l)|} \quad (21)$$

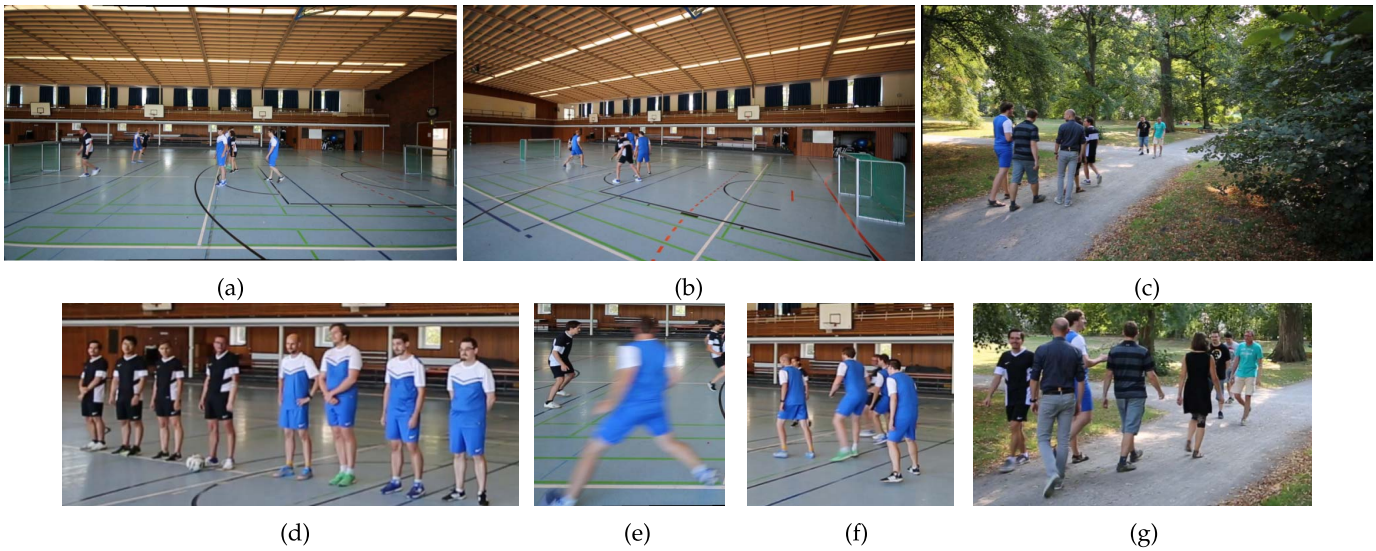

Fig. 5. First row: Different camera views and scenes in the VIMPT dataset. Second row: Challenges in the VIMPT dataset. (d) Very similar person appearances. (e) Rapid motions and motion blur. (f) Heavy occlusions. (g) Outdoor scene with frequent occlusions.

measures the mean squared distance between the acceleration signal $\mathbf{a}$ given by the IMU signal and the approximated acceleration $\hat{\mathbf{a}}$ derived from the video.

Thereby, the acceleration $\hat{\mathbf{a}}(\mathbf{p}, t)$ of a trajectory at time t is approximated via finite differences:

$$\hat{\mathbf{a}}(\mathbf{p}, t) := \frac{\mathbf{p}(t - \triangle t_{\text{IMU}}) - 2\mathbf{p}(t) + \mathbf{p}(t + \triangle t_{\text{IMU}})}{(\triangle t_{\text{IMU}})^2}, \quad (22)$$

where $\triangle t_{\text{IMU}}$ is the time distance between consecutive IMU signals. The set $T_{\text{IMU}}(l)$ denotes the timestamps within the first and last detection of $\mathcal{T}$ at which IMU signals exist.

The parameter w can be used to balance the importance of each input channel. The optimization problem (19) has a non-linear least squares form, and we apply the Levenberg-Marquardt algorithm [66], [67] to obtain a local optimum.

IV. Evaluation

In order to assess our proposed method, we recorded new sequences, since no dataset for the VIMPT setting exists so far. Our recordings contain challenging sequences captured with a calibrated camera and body-worn IMUs. An introduction of the dataset and details about the recording procedure are provided in Section IV-A. Further properties and challenges of the dataset are discussed in Section IV-B. In Section IV-C, we provide technical details of our tracking approach and assess its performance in Section IV-D. 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 state-of-the-art baselines.

A. VIMPT2019 Dataset

Existing benchmarks for video based people tracking do not contain IMU data. Hence, in order to evaluate our approach, we recorded a new dataset which we denote the VIMPT2019 dataset.

1) 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.

During soccer recordings, two four-person teams in team jerseys (see Figure 5(a,b,d)) play soccer in a competitive manner. Consequently, these recordings contain a lot of motion, motion blur, abrupt changes in direction, and occlusions, see Figure 5(e)-(f). Hence, tracking challenges arise from non-linear motion and ambiguous appearance information. Furthermore, 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 5(c,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 [22] and DukeMTMC [21], 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.

2) Camera Setup: For all sequences, a calibrated camera has been mounted to a tripod at a height of approximately 1.8m (see also Figure 5(a)). The videos were captured in landscape at 30Hz with 1920×1080 spatial resolution, and the camera's extrinsic matrix was calibrated to a fixed reference point in the scene.

3) Time Synchronization: All wireless IMU devices are automatically synchronized using the recording system of the IMU manufacturer. For the time synchronization between the IMU devices and the video, an additional IMU has been attached on a clapperboard. The clapperboard allows to detect the shut of the clap within the video and the IMU signal, respectively (see Figure 6).

4) Detections: We used the person detector Faster R-CNN [68] trained on COCO [69] to generate person detections

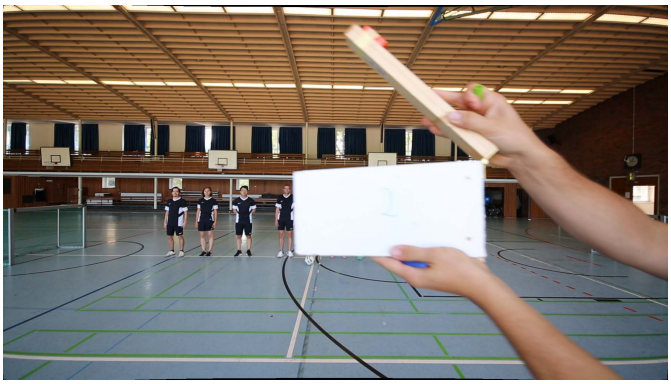

Fig. 6. Time synchronization using a clapper board.

TABLE I
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

TABLE II
AVERAGE ORIENTATION VARIANCE FOR ALL FOOTBALL SEQUENCES AND THE PARK SEQUENCE. HERE, σ_x^2 AND σ_z^2 DENOTE THE VARIANCE OF THE HEADING VECTOR IN x- AND z-DIRECTION IN GLOBAL COORDINATES, RESPECTIVELY

Name	σ_x^2	σ_z^2
Football	0.29	0.39
Walking	0.63	0.25

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 [22], we interpolated ground-truth detections for occluded persons.

5) 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.

6) Training, Validation and Test Split: We split the VIMPT dataset 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.

B. Characteristics of the VIMPT2019 Dataset

Many different state-of-the-art MPT datasets exist, each focusing on certain challenges of multiple people tracking. Some concentrate on a wide variety in the camera views [22], containing low- to semi-crowded scenes, while others focus on long-term tracking [21] with sometimes very crowded scenes, filmed from multiple static cameras. Our recordings focus on ambiguous appearance information and non-linear motions. Further, they deal as a proof-of-concept for the VIMPT setting.

1) Quantitative Characteristics: In this section, we provide further details of the VIMPT2019 dataset. Table I provides an overview of several recording characteristics. The lengths of the sequences range from 540 to nearly 1200 frames, which is similar to the MOT16/17 sequences. The third column of Table I 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. Ground-truth labels have been created by manually annotating ground-truth boxes. Note that people have also been labeled in situations of full occlusion using interpolation. We further provide a visibility distribution of the dataset. To this end, we compute for each ground-truth box $\mathbf{d}$ a visibility score $\text{vis}(\mathbf{d})$. 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 box foot position of $\mathbf{d}'$ is lower than the box foot position of $\mathbf{d}$. Then, we compute the relative number of pixels within $\mathbf{d}$ that are not occluded by other ground truth boxes:

$$\text{vis}(\mathbf{d}) = 1 - \frac{\left| \bigcup_{\mathbf{d} \lhd_C \mathbf{d}'} \mathbb{I}(\mathbf{d}') \cap \mathbb{I}(\mathbf{d}) \right|}{|\mathbb{I}(\mathbf{d})|}, \tag{23}$$

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 I and ranges between 0.79 and 0.84 with a standard deviation $\sigma^2(vis)$ between 0.10 and 0.19.

The concept of the VIMPT setting is to exploit local motion measurements from IMU sensors in order to compensate for ambiguous appearance or motion information. On the other hand, those local measurements may be very similar at a given timestep between people showing group behavior with similar walking aims (see e.g. [37]). We thus analyze the distribution of the orientations. To this end, we consider all heading directions $\mathbf{n} \in \mathbb{R}^2$ from the IMU sensors at a time step t to obtain the population variance $\sigma^2(t)$ of the orientations. Averaging over all frames and all sequences, we obtain the mean orientation variance σ_x^2 and σ_z^2 with respect to the x-axis and z-axis in the global coordinate frame. The results are presented in Table II. Thereby, moving in x-axis in global coordinates corresponds to moving closer or further to the camera and moving in z-direction corresponds to moving to the left or right.

Comparing the soccer sequences with the park sequence, the evaluations show that the orientations in the park sequence have a higher average variance. Accordingly, it is hard to distinguish the people of the soccer sequence directly from the orientation. Yet, as the subsequent experiments show, by taking all time steps of the sequence into account, our globally optimizing tracking system is able to resolve local ambiguities, thereby improving the tracking accuracy considerably.

2) Difficult Appearance: We provide a measurement of the appearance difficulty by employing standard re-identification metrics as a proxy. In particular, we compute the Top-1 and the mAP scores for the MOT16 dataset

TABLE III

DIFFICULTY OF RE-IDENTIFYING PERSONS IN TERMS OF THE TOP-1 AND MAP METRICS FOR THE MOT16 AND THE VIMPT2019 DATASET. IN ADDITION, WE ALSO PROVIDE THE METRICS EVALUATED ONLY ON THE FOOTBALL SEQUENCES (VIMPT2019*)

Dataset	Top-1	mAP
MOT16	90.3%	88.0%
VIMPT2019	67.4%	81.1%
VIMPT2019*	63.4%	78.3%

TABLE IV

COMPARISON OF THE VELOCITY DISTRIBUTIONS IN TERMS OF THE MEAN VELOCITY $\overline{v}$ AND ITS VARIANCE $\sigma^2(v)$ FOR THE MOT16 AND THE VIMPT2019 DATASETS. IN ADDITION WE ALSO PROVIDE THE METRICS EVALUATED ONLY ON THE FOOTBALL SEQUENCES (VIMPT2019*)

Dataset	$\overline{v}$	$\sigma^2(v)$
MOT16	0.19	0.06
VIMPT2019	0.60	0.34
VIMPT2019*	0.63	0.36

and the VIMPT2019 dataset using the state-of-the-art re-identification system [70] fine-tuned on DukeMTMC [21]. Further, 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 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 III, 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 are ambiguous due to the soccer jerseys.

3) Motion Characterizations: We characterize the velocity distribution of the VIMPT2019 dataset. To this end, we use the ground-truth trajectories and compute the average image velocity between two detections $\mathbf{d}_1$ and $\mathbf{d}_2$ belonging to the same person and normalize it by the mean box height of $\mathbf{d}_1$ and $\mathbf{d}_2$, in order to compensate perspective effects. To put the values into perspective, we compare to the MOT16 dataset[2] and present the results in Table IV. The evaluations show that VIMPT2019 contains much higher velocities, and most importantly, the velocities 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.

C. Tracker Parameters

1) Tracklet Generation: We generate reliable tracklets by grouping detections using the method of [71]. 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.

2) Visual Heading Network: The overall network architecture is depicted in Figure 4. It contains the VGG16 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 VGG16 has been trained on ImageNet with an invariance for horizontal flipping [62]. To undo this, we train the layers FC1, FC2, and FC3 together with the last convolutional layer of VGG16 while keeping the weights of all other layers fixed. During training, we add dropout layers [72] with $p = 0.3$ between the fully connected layers to avoid overfitting. Furthermore, 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 [73] 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

$$\frac{1}{|D|} \sum_{\mathbf{d} \in D} \Phi(\hat{\mathbf{n}}_{\mathbf{d}}(\mathbf{W}), \mathbf{n}_{\mathbf{d}}), \tag{24}$$

given all ground-truth detections $\mathbf{d} \in D$ and corresponding IMU heading vectors $\mathbf{n}_{\mathbf{d}}$ of the VIMPT training sequence.

3) Graph Edge Settings: In the graph $\mathcal{G}$, weighted edges $e \in \mathcal{E}$ are created between two nodes v and v' in the following cases. If the shortest temporal distance between all detections of v and v' is at most 12 frames, we establish a short-term edge associated with costs derived from box features. Similarly, we establish long-term edges associated with costs derived from acceleration features between all detections of v and v' if the temporal distance is between 12 and 150 frames.

4) Feature to Cost Mapping: In order to transform unary and pairwise features to costs, we use different strategies. For orientation and box features we apply a logistic regression model [74]. A feature vector $\boldsymbol{f}$ is mapped to costs $c = -\langle \mathbf{f}, \mathbf{w} \rangle$, by learning the appropriate weight $\mathbf{w}$, so that the optimization problem (1) is probabilistically motivated [75]. We use ground-truth trajectories in the training sequence of the VIMPT dataset to train the model parameters $\mathbf{w}$. This does not work satisfactorily for the acceleration feature. We observed that noise in 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 if $\min f_{\text{acc}}(\mathcal{H}, \mathcal{H}') > \delta$.

D. Tracking Evaluation

The goal of this work is to track IMU-equipped persons in a video accurately. Hence a perfect tracking result is achieved if the assignment of person-specific IDs to corresponding tracklets is coherent throughout the whole tracking sequence.

1) Error Metrics: We evaluate tracking performance by assessing assignment coherency in terms of ID metrics. According to [21] we compute IDP, IDR and IDF1. IDP is the ID precision measuring the fraction of ground-truth person detections that are correctly assigned to a unique person ID. Similarly, IDR is the recall rate of respective ground-truth

[2]We use all training sequences which are filmed from a static camera.

TABLE V
TRACKING ACCURACY SOCCER SEQUENCES

Tracker	IDP ↑	IDR ↑	IDs ↓	MOTA ↑	IDF1 ↑
DeepCC [14]	26.3	27.9	395	11.8	27.1
DeepSORT [40]	49.6	42.4	193	77.1	45.8
FWT [13]	29.7	26.7	489	71.6	28.1
VIT	**93.6**	**90.1**	**44**	**86.1**	**91.8**

TABLE VI
TRACKING ACCURACY OUTDOOR RECORDING

Tracker	IDP ↑	IDR ↑	IDs ↓	MOTA ↑	IDF1 ↑
DeepCC [14]	55.4	57.0	47	67.3	56.2
DeepSORT [40]	53.4	48.0	28	**83.5**	50.5
FWT [13]	39.1	36.3	66	82.4	37.6
VIT	**89.5**	**78.5**	**22**	81.8	**88.5**

detections. The metric IDF1 is the ratio of correctly identified detections over the average number of ground-truth and computed detections. The basic idea of IDF1 is to combine IDP and IDR to a single number.

In addition to the aforementioned ID metrics, we report the CLEAR-MOT metric MOTA [76]. MOTA comprises three different error metrics, namely the number of ID switches, false positives, and false negatives. Note that the computations of false positives and false negatives within MOTA is based solely on detection existence. Hence, detections from computed trajectories are matched to ground truth detections for each frame separately, ignoring any ID consistency checks. In constrast to that, IDF1 evaluates false positives (negatives) *and* also verifies that the person ID is correct. In particular, the IDF1 score incorporates the longest coverage of each ground 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.

2) Tracking Accuracy: We report tracking accuracy of our approach, denoted as Video Inertial Tracker (VIT), on the VIMPT dataset in the bottom rows of Table V and Table VI. 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 the 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 sport tracking but generalizes to other scenarios too. Note that the Faster-RCNN 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. In the supplementary material, we provide a video showing the sequences of the dataset and corresponding tracking results.

3) Comparison to Vision-Based Methods: We apply three different state-of-the-art vision-based trackers to the VIMPT dataset, namely FWT [13], DeepSORT [39], and DeepCC [14]. FWT is among the top performing trackers of the MOT17 [22] benchmark, 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 appearance information is very ambiguous. The tracking results shown in Table V validate that this is very challenging for all considered state-of-the-art 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 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. Further note that no competing multiple people tracking method exists which fuses video and IMU information.

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 V with Table VI 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 the misleading information.

4) Effectiveness of Interpolation: We analyze the benefit of having IMU information available in terms of reconstructing missing detections. To this end, we compute the tracking performance using the interpolation method introduced in Section III-E and analyze its robustness for all sequences of the VIMPT2019 dataset. Further, 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 7. The experiment shows significantly better performance when the interpolation 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. The improvement over video-based interpolation increases significantly as more detections are removed.

Accordingly, using the IMU interpolation, the tracking approach becomes more robust against occlusions and less dependent on appearance information. When 10% of the input detections are removed, the performance using the

IMU interpolation 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 80% for the IMU interpolation, while it drops to about 30% using the linear interpolation. This experiment clearly demonstrates the benefits of the VIMPT setting. Objects of interests may be occluded or even out of view and can still be tracked reliably.

5) 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 VII.

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, respectively. 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 successfully differentiate people. This is due to the soccer sequences, where soccer players are often oriented similarly to follow the game ball (see also Table II). Consequently, these sequences provoke a high number of IDS. In this sense, the soccer recordings can be seen as a worst-case test setup.

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 interpolation method based on the IMU's acceleration signal. This helps to recover more detections and to form consistent trajectories, which are closer to ground-truth in the end. 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 weakend 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 are key to disambiguate tracklet assignments and help to correctly reject most of implausible hypotheses. The tracker Ori+Acc is leveraging the IMU signals for the features and for the interpolation. The video information is used only to compare orientations in the video and the IMU signal. 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 Ori+Acc is independent of any artifical motion model or of the constancy assumptions on the appearance, as it just compares measurements. This shows the potential of the VIMPT setting.

TABLE VII

TRACKING ACCURACY OF FOUR TRACKER VARIANTS AND OUR PROPOSED TRACKER (VIT), EVALUATED ON ALL SEQUENCES

Tracker	IDP ↑	IDR ↑	IDs ↓	MOTA ↑	IDF1 ↑
Ori	48.1	49.3	1083	45.1	48.7
VT	38.0	38.1	267	52.0	38.1
VT+Acc	44.9	45.1	256	65.0	45.0
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**

TABLE VIII

TRAINING AND TEST ACCURACY OF THE VISUAL HEADING NETWORK. WE PROVIDE THE RELATIVE NUMBER OF HEADING ERRORS WITHIN A THRESHOLD OF $\epsilon \in \{30°, 45°\}$

	$\leq 45°$	$\leq 30°$
Train[PC]	**97.2**%	**88.8**%
Train[w/o PC]	96.7%	87.5%
Test[PC]	**96.2**%	**88.1**%
Test[w/o PC]	86.3%	70.4%

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. Finally, Ori shows the results when all pairwise costs are set to zero, thereby relying completely on the VHN network.

6) Visual Heading Network Accuracy: We evaluate the Visual Heading Network accuracy by computing the relative number of predicted heading vectors $\hat{\mathbf{n}}_\mathbf{d}$ that deviate not more than ϵ degrees from ground-truth. The network is trained on the VIMPT training sequence and tested on all other sequences of the dataset. According to Table VIII, the network predicts orientations with high accuracy and is able to generalize to unseen images (Train[PC] and Test[PC]). Also note that the perspective correction 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 not able to generalize the observed perspectives from 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.

7) Runtime: In general, solving binary quadratic problems such as (1) is very challenging. However, in our experiments we observed very fast solutions. On a Intel i9 CPU with 8 cores and 3.60GHz, the runtime of the solver[3] 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 (1) increases with the number of people to be tracked. If runtime becomes critical, various approximative solvers [13], [77], [78] could be applied to accelerate the computations.

8) Identification Accuracy: According to [21] the ID precision metric (IDP) evaluates if all tracklets of a person are correctly assigned to a unique ID $i \in \mathbb{N}$. However, this does not necessarily mean that a person's trajectory is assigned to the person label $j \in \mathcal{L}$ defined by the corresponding IMU

[3]we used gurobi in version 9.02.

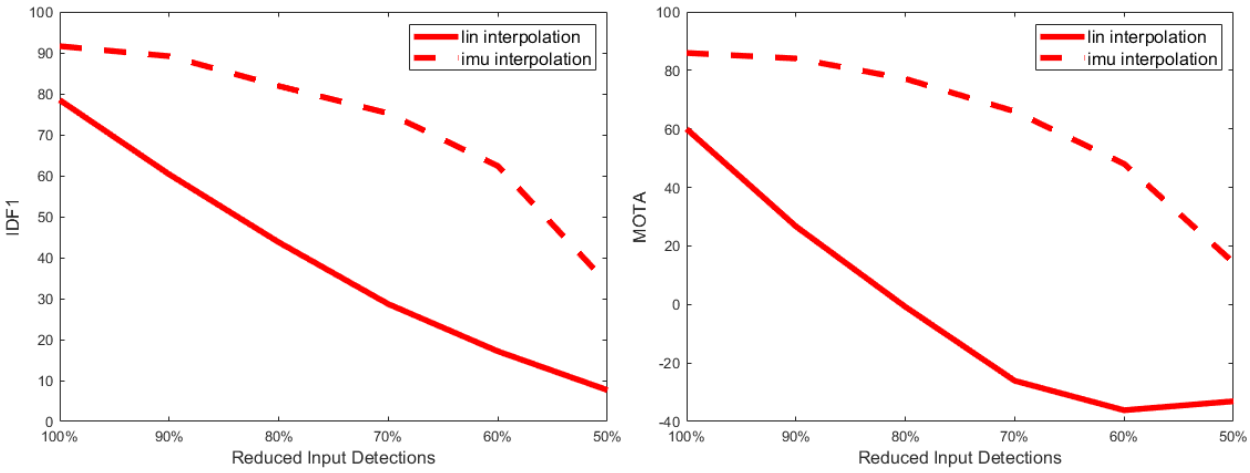

Fig. 7. Impact of different interpolation methods, when detections are missing.

device. Hence, we manually investigated if each ID i actually corresponds to the associated IMU ID j. This is the case for all persons and sequences in the VIMPT dataset.

Within the VIMPT setting, our method is thus able to simultaneously track and identify IMU equipped people from a video.

V. Conclusion and Future Work

This work introduces a novel extension to the common multiple people tracking problem; combining video information with measurements from body-worn IMUs for the purpose of multiple people tracking, which we call Video Inertial Multiple People Tracking (VIMPT). Conceptually, this setting 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 the correspondence of objects and IMUs is known in advance, a tracking solution automatically provides object identities within the video. To tackle the challenging VIMPT problem, we proposed a graph labeling formulation to assign tracklets in the video to corresponding IMU devices, such that the assignments are consistent with the video information and the IMU signals at the same time. The IMU orientations are correlated to predicted orientations of the corresponding objects in the video. For this purpose, we propose a neural network with a novel perspective correction procedure, which turned out to be essential for accurate video to IMU assignments. Accelerations measured from the IMU sensors were correlated to initial velocities, as measured in the video. Besides the proposed tracking approach, this work releases the first VIMPT dataset, which we called VIMPT2019. We use this dataset to evaluate the effectiveness of the VIMPT setting and to benchmark our proposed tracker. 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 where 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 which is independent of person appearance and still able to track fast and dynamic motions. In the experiments, we validate that our proposed tracker shows substantial improvements over video-based tracking systems. This demonstrates the potential of the VIMPT setting. Yet, the proposed tracker 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. As future work, we plan to extend our approach to work with an uncalibrated camera. In order to improve orientation predictions from the video, we plan to integrate attention mechanisms (in the spirit of [45]) that estimate the orientation only for those detections of a tracklet that do not contain impaired visual information, such as partial occlusions or motion blur. We further plan to transform our proposed tracker to an end-to-end trainable tracking system, inspired by the current progress in this direction [79], [80] for other tracking systems. While we demonstrated that a fusion of Video data with IMU signals improves multiple people tracking systems, the same concept could be applied to track other objects, which would extend our setup to VIMOT (Video Inertial Mulit-Object Tracking). Thus, another future direction of research is to apply the VIMOT setting to various other scenarios, especially for cases where it is very hard to distinguish individual objects. One application in mind is tracking and identifying individual animals in a herd.

References

[1] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3488–3496.

[2] A. Alahi, V. Ramanathan, and L. Fei-Fei, "Socially-aware large-scale crowd forecasting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2203–2210.

[3] A. Alahi *et al.*, "Learning to predict human behavior in crowded scenes," in *Group and Crowd Behavior for Computer Vision*. Amsterdam, The Netherlands: Elsevier, 2017, pp. 183–207.

[4] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971

[5] A. Alahi, J. Wilson, L. Fei-Fei, and S. Savarese, "Unsupervised camera localization in crowded spaces," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 2666–2673.

[6] M. Fenzi *et al.*, "ASEV—Automatic situation assessment for event-driven video analysis," in *Proc. 11th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2014, pp. 37–43.

[7] A. Alahi, Y. Boursier, L. Jacques, and P. Vandergheynst, "Sport players detection and tracking with a mixed network of planar and omnidirectional cameras," in *Proc. 3rd ACM/IEEE Int. Conf. Distrib. Smart Cameras (ICDSC)*, Aug. 2009, pp. 1–8.

[8] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1704–1716, Jul. 2013.

[9] R. Li and B. Bhanu, "Fine-grained visual dribbling style analysis for soccer videos with augmented dribble energy image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–9.

[10] R. Theagarajan, F. Pala, X. Zhang, and B. Bhanu, "Soccer: Who has the ball? Generating visual analytics and player statistics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1749–1757.

[11] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn, "Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 120–127.

[12] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Multi-person tracking by multicut and deep matching," in *Proc. ECCV Workshop Benchmarking Multi-Target Tracking (ECCVW)*. Cham, Switzerland: Springer, 2016, pp. 100–111.

[13] R. Henschel, L. Leal-Taixe, D. Cremers, and B. Rosenhahn, "Fusion of head and full-body detectors for multi-object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1428–1437.

[14] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6036–6046.

[15] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3539–3548.

[16] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," 2018, *arXiv:1812.02162*. [Online]. Available: https://arxiv.org/abs/1812.02162

[17] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-Identification*. Cham, Switzerland: Springer, 2014.

[18] A. R. Jimenez, F. Seco, C. Prieto, and J. Guevara, "A comparison of pedestrian dead-reckoning algorithms using a low-cost MEMS IMU," in *Proc. IEEE Int. Symp. Intell. Signal Process.*, Aug. 2009, pp. 37–42.

[19] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse inertial poser: Automatic 3d human pose estimation from sparse IMUs," *Proc. 38th Annu. Conf. Eur. Assoc. Comput. Graph. (Eurographics)*, 2017, pp. 349–360.

[20] S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[21] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. ECCV Workshop Benchmarking Multi-Target Tracking (ECCVW)*, 2016, pp. 17–35.

[22] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," 2016, *arXiv:1603.00831*. [Online]. Available: http://arxiv.org/abs/1603.00831

[23] R. Henschel, T. von Marcard, and B. Rosenhahn, "Simultaneous identification and tracking of multiple people using video and IMUs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[24] R. Henschel, L. Leal-Taixé, and B. Rosenhahn, "Efficient multiple people tracking using minimum cost arborescences," in *Proc. German Conf. Pattern Recognit. (GCPR)*, 2014, pp. 265–276.

[25] R. Henschel, L. Leal-Taixé, B. Rosenhahn, and K. Schindler, "Tracking with multi-level features," 2016, *arXiv:1607.07304*. [Online]. Available: http://arxiv.org/abs/1607.07304

[26] W. Choi, "Near-online multi-target tracking with aggregated local flow descriptor," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3029–3037.

[27] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele, "Motion segmentation & multiple object tracking by correlation co-clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 140–153, Jan. 2020.

[28] E. Levinkov *et al.*, "Joint graph decomposition & node labeling: Problem, algorithms, applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1904–1912.

[29] A. Dehghan, S. M. Assari, and M. Shah, "GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4091–4099.

[30] P. Chu and H. Ling, "FAMNet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking," 2019, *arXiv:1904.04989*. [Online]. Available: http://arxiv.org/abs/1904.04989

[31] P. Bergmann, T. Meinhardt, and L. Leal-Taixe, "Tracking without bells and whistles," 2019, *arXiv:1903.05625*. [Online]. Available: http://arxiv.org/abs/1903.05625

[32] A. Hornakova, R. Henschel, B. Rosenhahn, and P. Swoboda, "Lifted disjoint paths with application in multiple object tracking," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, Jul. 2020, pp. 1–22.

[33] B. Yang and R. Nevatia, "Multi-target tracking by online learning of non-linear motion patterns and robust appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1918–1925.

[34] J. Shen, Z. Liang, J. Liu, H. Sun, L. Shao, and D. Tao, "Multiobject tracking by submodular optimization," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 1990–2001, Jun. 2019.

[35] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[36] J. Shen, J. Peng, and L. Shao, "Submodular trajectories for better motion segmentation in videos," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2688–2700, Jun. 2018.

[37] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 261–268.

[38] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3464–3468.

[39] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 748–756.

[40] L. Leal-Taixe, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, "Learning an image-based motion context for multiple people tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3542–3549.

[41] C.-H. Kuo and R. Nevatia, "How does person identity recognition help multi-person tracking?" in *Proc. CVPR*, Jun. 2011, pp. 1217–1224.

[42] C. Kim, F. Li, and J. M. Rehg, "Multi-object tracking with neural gating using bilinear LSTM," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 200–215.

[43] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang, "Online multi-object tracking with dual matching attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 366–382.

[44] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4836–4845.

[45] J. Xu, Y. Cao, Z. Zhang, and H. Hu, "Spatial-temporal relation networks for multi-object tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3988–3998.

[46] X. Dong, J. Shen, W. Wang, L. Shao, H. Ling, and F. Porikli, "Dynamical hyperparameter optimization via deep reinforcement learning in tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 29, 2019, doi: 10.1109/TPAMI.2019.2956703.

[47] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3516–3527, Jul. 2019.

[48] Z. Liang and J. Shen, "Local semantic siamese networks for fast tracking," *IEEE Trans. Image Process.*, vol. 29, pp. 3351–3364, 2020.

[49] E. S. Jones and S. Soatto, "Visual-inertial navigation, mapping and localization: A scalable real-time causal approach," *Int. J. Robot. Res.*, vol. 30, no. 4, pp. 407–430, Apr. 2011.

[50] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," Georgia Inst. Technol., Atlanta, GA, USA, Tech. Rep., 2015.
[51] T. V. Marcard, G. Pons-Moll, and B. Rosenhahn, "Human pose estimation from video and IMUs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1533–1547, Aug. 2016.
[52] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 601–617.
[53] J. Yan, G. He, A. Basiri, and C. Hancock, "3-D passive-vision-aided pedestrian dead reckoning for indoor positioning," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 4, pp. 1370–1386, Apr. 2020.
[54] W. Jiang and Z. Yin, "Combining passive visual cameras and active imu sensors to track cooperative people," in *Proc. Int. Conf. Inf. Fusion*, 2015, pp. 1338–1345.
[55] W. Jiang and Z. Yin, "Combining passive visual cameras and active IMU sensors for persistent pedestrian tracking," *J. Vis. Commun. Image Represent.*, vol. 48, pp. 419–431, Oct. 2017.
[56] M. Camplani *et al.*, "Multiple human tracking in RGB-depth data: A survey," *IET Comput. Vis.*, vol. 11, no. 4, pp. 265–285, Jun. 2017.
[57] A. Alahi, A. Haque, and L. Fei-Fei, "RGB-W: When vision meets wireless," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3289–3297.
[58] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3271–3279.
[59] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2010, pp. 282–295.
[60] P. Nillius, J. Sullivan, and S. Carlsson, "Multi-target tracking–linking identities using Bayesian network inference," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. New York, NY, USA: IEEE Computer Society, Jun. 2006, pp. 2187–2194.
[61] M. Kok, J. D. Hol, and T. B. Schön, "Using inertial sensors for position and orientation estimation," *Found. Trends Signal Process.*, vol. 11, nos. 1–2, pp. 1–153, 2017.
[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[63] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[64] L. J. Watters, "Reduction of integer polynomial programming problems to zero-one linear programming problems," *Oper. Res.*, vol. 15, no. 6, pp. 1171–1174, 1967. [Online]. Available: http://www.jstor.org/stable/168623
[65] Gurobi Optimization. (2018). *Gurobi Optimizer Reference Manual*. [Online]. Available: http://www.gurobi.com
[66] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quart. Appl. Math.*, vol. 2, no. 2, pp. 164–168, 1944.
[67] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, no. 2, pp. 431–441, Jun. 1963.
[68] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.
[69] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.
[70] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 402–419.
[71] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects," in *Proc. CVPR*, Jun. 2011, pp. 1201–1208.
[72] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[73] T. Tieleman and G. Hinton, "Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude," *COURSERA, Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.
[74] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.
[75] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Subgraph decomposition for multi-target tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5033–5041.
[76] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The CLEAR MOT metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, Dec. 2008.
[77] J. Shen, J. Peng, X. Dong, L. Shao, and F. Porikli, "Higher order energies for image segmentation," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4911–4922, Oct. 2017.
[78] X. Dong, J. Shen, L. Shao, and L. Van Gool, "Sub-Markov random walk for image segmentation," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 516–527, Feb. 2016.
[79] S. Schulter, P. Vernaza, W. Choi, and M. Chandraker, "Deep network flow for multi-object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6951–6960.
[80] J. Xiang, G. Xu, C. Ma, and J. Hou, "End-to-end learning deep CRF models for multi-object tracking," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Feb. 24, 2020, doi: 10.1109/TCSVT.2020.2975842.

Roberto Henschel received the Dipl.-Math. degree from the Free University of Berlin in 2012. He studied mathematics with a minor in computer science. He is currently pursuing the Ph.D. degree with the Leibniz University of Hannover. From 2012 to 2013, he was with Fachbereich Mathematik, TU Darmstadt, as a Research Assistant. His research interests include multiple people tracking and sensor fusion. His works have led to the second place in the WACV 2015 Multi-Target Tracking Challenge and the first place in the CVPR 2017 Multi Object Tracking Challenge.

Timo von Marcard received the Dipl.Ing.(FH) degree in mechatronics from the University of Applied Sciences Karlsruhe in 2008 and the M.Sc. degree in systems, control and mechatronics from the Chalmers University of Technology, Gothenburg, in 2010. He was a Developer for embedded software and algorithms at Otto Bock Healthcare GmbH. His research interests include marker-less human motion capture and sensor fusion algorithms. His work *Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs* received the Günter Enderle Award for the best article at Eurographics 2017.

Bodo Rosenhahn received the Dipl.-Inf. and Dr. Ing. degrees from the University of Kiel in 1999 and 2003, respectively. He studied computer science (minor in subject medicine) with the University of Kiel. From 2003 to 2005, he held a (DFG) postdoctoral position with The University of Auckland, New Zealand. From 2005 to 2008, he was a Senior Researcher with the Max-Planck Institute for Computer Science, Saarbruecken, Germany. Since 2008, he has been a Full Professor with the Leibniz University of Hannover, heading a group on automated image interpretation. He has written more than 180 research articles, holds more than ten patents, and received several awards. His research interests include computer vision and machine learning.