Simultaneous Identification and Tracking of Multiple People using Video and IMUs

Roberto Henschel Timo von Marcard Bodo Rosenhahn
Leibniz Universität Hannover, Germany
{henschel,marcard,rosenhahn}@tnt.uni-hannover.de

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 to deal 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.

1. 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 [48, 4, 5, 2], support urban planning [6], secure areas against dangerous behavior [15] or to provide an automatic analysis of player's performance in sports [1, 31].

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 formulated, 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 whether two detections belong to the same person or not. Typically, this involves a motion model or person appearance. A motion model attempts to assign likelihoods to observed person movements. This is very generic and only depends on corner coordinates of detection boxes. However, as soon as the motion becomes more dynamic, simple motion models [28] are insufficient and tracking degrades. In particular, most motion models assume low and constant velocities, which holds for pedestrians only within a short temporal window [40].

Another complementary strategy is to model relations between detections based on person appearance. Here, CNN-based feature representations are used to evaluate if two detections show the same person. Recent works have shown very impressive tracking results

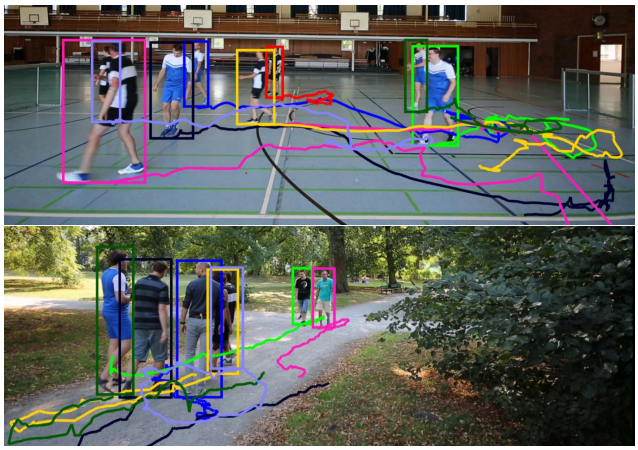

Figure 1: Qualitative results obtained by fusing person detections and local motion measurements of body-worn IMUs. Instead of relying on person appearance, the proposed approach enables accurate long-term tracking by finding a globally optimal assignment of detection boxes to IMU devices, such that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

using this information exclusively [18, 40] or in combination with motion models [37, 41]. A major advantage of utilizing appearance information over motion models is that they allow to relate detections which are temporally far apart. This facilitates to re-identify people even after long-term occlusions or if they temporally fall out of the camera view.

Despite the enormous progress with Neural Network-based appearance features, it remains challenging to differentiate persons wearing similar or identical clothing. A prototypical example for such a situation is sport player tracking, where team members wear almost identical dresses. Another challenge arises 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 consequently tracking accuracy degrades.

In this work we propose an alternative tracking method that does not rely on appearance features and is still capable to deal with very dynamic motions and heavy occlusions. We got inspired by other works in the field of human motion capture [43, 42] and SLAM [23]. In these fields vision has been combined with additional sensor modalities provided from inertial measurement units (IMUs). IMUs are small motion sensors measuring local orientation and acceleration.

Incorporating additional sensory input for the task of MPT creates a very different problem setup compared to the aforementioned vision-only methods. Hence, we will refer to this setting as Video Inertial Multiple People Tracking (VIMPT).

In VIMPT 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 assignment 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 to track and re-identify 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 night-vision is used or for a sports team during training. In summary: VIMPT allows to track people and at the same time to recover their identities in terms of the associated IMU devices.

Even though in VIMPT motion information is available through IMU measurements it 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 [22, 44]. If this was possible, we 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 by its own. In particular, this requires to relate IMU orientations, which are elements of SO(3), 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.

In this work we propose a new method that can cope with the aforementioned challenges. In particular, the method enables long-term tracking of multiple people without using person appearance, see Figure 1. We make this feasible by

- integrating inertial measurements from body-worn IMUs,
- designing a neural network to relate person detections to orientation measurements,
- finding a globally coherent assignment of IMU devices to person detections, by integrating all available information in a single graph labeling formulation.

In order to evaluate our proposed method, we recorded a new dataset containing challenging soccer sequences and a regular outdoor scene. We demonstrate that our approach is capable to accurately track and identify persons during fast and dynamic motions. This even works reliably under heavy occlusions and if they temporally leave the field of view.

To the best of our knowledge, we are the first who achieve a fully automatic system that integrates IMUs into a multi-people tracking framework in order to improve the accuracy and to obtain automatic person assignments.

2. Related Work

Data Association. Most multiple people tracking works employ the tracking-by-detection paradigm [18, 41, 19, 20, 10, 24, 37, 29, 12] that connects either detections [18, 41, 29] or precomputed tracklets [47, 10, 12] to form the trajectories. The problem of creating trajectories is usually formulated as a graph optimization problem. Several works apply network-flow [28, 49], while more recently minimum cost multicut [41, 24, 40] or graph labeling [18] formulations have been proposed.

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 [28, 33, 47, 8, 46, 27]. Typically a linear constant velocity model within short time windows is assumed [28, 8]. However, the performance of these approaches degrades if motions become more dynamic or people get temporally occluded. Consequently, current state-of-the-art tracking systems [26, 41, 40, 24, 37, 18, 25, 50, 11] rely on appearance models which are invariant to these issues. They use sophisticated neural networks to derive association weights from person appearance. These association weights have improved to a level that some works reformulate the tracking problem as a person re-identification problem [11, 37].

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 lightning variations can change the perceived appearance of a person.

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

Vision and Inertial Sensors. Body-worn inertial sensors provide motion information independent of person visibility. However, it is not possible to recover the 3D person trajectory from IMU information alone [22, 44]. In contrast, video allows to extract positional information, which is complementary to the IMU motion information.

Consequently, IMUs have been combined with visual information in many application, e.g. the works [23, 16] fuse video and inertial data to stabilize self localization and mapping (SLAM). The same modalities have been used to recover human poses in [43, 42].

There exist only very few works that incorporate IMUs for people tracking in videos. The closest reference to our work is [21], which tackles single person tracking. An IMU-equipped person has to be manually localized in the first video frame. Then, IMU information is 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.

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. [9] provides a survey of tracking approaches using RGB-D cameras and [3] integrates wireless signals emitted from cell phones.

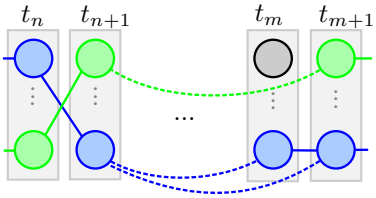

Figure 2: Every tracklet represents a node in the graph. Each node can be assigned to an IMU device (indicated by color) and is linked to other nodes by short-term edges (solid) and long-term edges (dashed). An edge is activated if corresponding nodes share the same color. The idea is that every graph color configuration is associated to costs representing consistency of video information and IMU data. The goal is to find the assignment with minimal costs.

3. Method

We follow the tracking-by-detection paradigm and group detections to short tracklets in a first step.

Then the tracking task can 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 want solve the tracking task 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 to 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. Note that using initial tracklets allows (i) to reduce the problem size and (ii) to make the orientation regression more robust to occasional cases where multiple people appear within a detection box.

3.1. Model

In order to solve the tracking task, we create 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 set $\mathcal{L} = \{1, 2, 3, ..., P\}$ contains an ID for all P persons wearing an IMU.

We introduce the notion of an assignment hypothesis $\mathcal{H} = (v, l)$, which associates a label $l \in \mathcal{L}$ to tracklet $v \in \mathcal{V}$. Associated to each hypothesis are assignment costs $c_v^l \in \mathcal{C}$ and indicator variables x_v^l, which take value 1 if $\mathcal{H}$ 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 casted into a binary optimization problem:

$$\underset{x \in \mathcal{F} \cap \{0,1\}^{|\mathcal{V}|P}}{\arg\min} \sum_{l \in \{1, \cdots, P\}} \left(\sum_{v \in \mathcal{V}} c_v^l x_v^l + \sum_{e \in \mathcal{E}} c_e^l \prod_{v \in e} x_v^l \right), \tag{1}$$

where the feasibility set $\mathcal{F}$ is subject to

$$\forall v \in \mathcal{V} : \sum_{l=1}^{P} x_v^l \leq 1, \tag{2}$$

$$\forall t, \forall l \in \{1, \ldots, P\} : \sum_{v \in \mathcal{V}_t} x_v^l \leq 1. \tag{3}$$

The subset $\mathcal{V}_t \subset \mathcal{V}$ comprises all tracklets v that contain a detection in frame t. Eq. (2) ensures that each tracklet v 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.

Next, we describe the unary and pairwise potentials in detail. Specifically, we introduce consistency features which are later mapped to costs c_v^l and c_e^l.

3.2. Unary Features

In order to provide a measure for the likelihood of an assignment hypothesis $\mathcal{H} = (v, l)$, we estimate the person orientation 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 torsos coronal plane projected to the ground plane as illustrated in Figure 3a. We use the projected normal as this comprises less degrees of freedom and people usually move in a rather upright pose.

Hence, given the image data I_d of detection d we seek to estimate the heading $\hat{\mathbf{n}}_d$ of the person. However, the observed heading in I_d depends on the person position

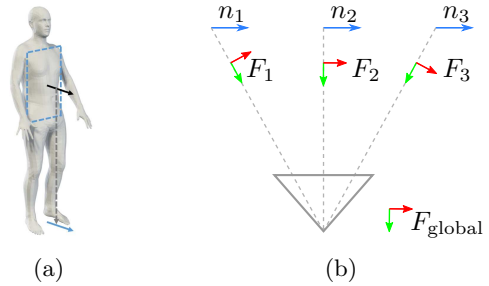

Figure 3: (a) We define person orientation in terms of the normal vector of the torso's coronal plane (black arrow) projected to the ground plane (blue arrow). (b) Even though global orientation (blue arrows) is constant at depicted positions, the perceived orientation as seen from the camera varies.

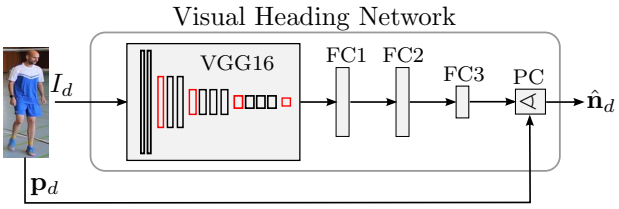

Figure 4: The Visual Heading Network predicts the heading $\hat{\mathbf{n}}_d$ of a person using the image data I_d of detection d. Based on the box position $\mathbf{p}_d$, the network performs a perspective correction (PC) in the last layer.

in the image, see Figure 3b. 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 view point 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 α_d be the angle between the vector defined by the camera center and box position $\mathbf{p}_d$, and the depth-axis of the camera. In order to compensate the perspective influence, we rotate the perceived orientation by $-\alpha_d$ and obtain the prediction $\hat{\mathbf{n}}_d$, cmp. Figure 3b.

In order to obtain the person heading from image data we employ a Neural Net to learn the mapping $I_d \mapsto \hat{\mathbf{n}}_d$. More specifically, we extend VGG16[38] pretrained on ImageNet[13] 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 (VHN) in the following. A graphical illustration showing the network architecture is depicted in Figure 4.

In the VIMPT setting, IMUs are consistently placed

at the back of each person such that the local sensor z-axis corresponds to the normal vector of the torsos coronal plane. Hence, we get the measured torso orientation vector $n_{l,t}$ 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 Π projects the normal vector to the ground plane.

Finally, we define the unary orientation feature representing the likelihood of hypothesis $\mathcal{H}$ as

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

where Φ denotes the cosine similarity, N_d corresponds to the number of detections of tracklet v and t_d represents the time stamp of a detection d.

3.3. 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.

Box 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. Accordingly, we employ well-established spatio-temporal features, which allow the tracker to be independent of appearance information. We derive corresponding features in the following.

For each detection box d we get a rough 3D position estimate $\mathbf{p}_d \in \mathbb{R}^3$ by projecting the detection box foot point to the 3D ground plane of the scene. Hence, for detections d of v and d' of v' let $\mathbf{v}_{\text{3D}}(d, d')$ denote the velocity in 3D from d to d'. Let $N(v, v')$ be the set of all pairs of detections between $\mathcal{H}$ and $\mathcal{H}'$ considered for the feature. We define the mean velocity feature between $\mathcal{H}$ and $\mathcal{H}'$ as

$$f_{\text{vel}}(\mathcal{H}, \mathcal{H}') = \frac{1}{|N(v, v')|}\sum_{(d,d')\in N(v,v')} ||\mathbf{v}_{\text{3D}}(d, d')||_2. \tag{6}$$

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

$$\Delta_h(d, d') = \nabla_t(d, d')\frac{|h_d - h_{d'}|}{\min\{h_d, h_{d'}\}}, \tag{7}$$

where the factor in front of the fraction compensates for the temporal distance between d and d':

$$\nabla_t(d, d') = \frac{1}{\log(2 + |t_d - t_{d'}|)}. \tag{8}$$

Finally, we define a box height feature as

$$f_{\text{height}}(\mathcal{H}, \mathcal{H}') = \frac{1}{|N(v, v')|}\sum_{(d,d')\in N(v,v')}\Delta_h(d, d'). \tag{9}$$

Both, f_{vel} and f_{height} 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 which cover larger temporal horizons.

Acceleration Feature. Ideally, the position $\mathbf{p}_{t_1} \in \mathbb{R}^3$ at time t_1 of an IMU can be recovered by double integration of the corresponding acceleration signal $\mathbf{a}$ according to

$$\mathbf{p}_{t_1} = \mathbf{p}_{t_0} + \mathbf{v}_{t_0}(t_1 - t_0) + \int_{t_0}^{t_1}\int_{t_0}^{u}\mathbf{a}(s)ds\,du, \tag{10}$$

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.

Let $\mathbf{p}_{t_0}$ be the 3D position of detection d and $\mathbf{p}_{t_1}$ the 3D position of d'. After double integration of the acceleration signal, we can solve Eq. (10) for the initial velocity, which we denote $\mathbf{v}_{\text{IMU}}(d, d')$. Concurrently we can approximate a persons velocity $\mathbf{v}_d$ at initial time t_0 in terms of finite differences of neighboring detections of d. Hence, for a compatible hypotheses pair $\mathcal{H}$ and $\mathcal{H}'$ the velocity differences

$$\Delta_v(d, d') = ||\mathbf{v}_{\text{IMU}}(d, d') - \mathbf{v}_d||_2 \tag{11}$$

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

$$f_{\text{acc}}(\mathcal{H}, \mathcal{H}') = \{\Delta_v(d, d')\,|\,(d, d') \in N(v, v')\}. \tag{12}$$

3.4. 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 binary variable $z_{v,v'}^l$ and the following constraints are added:

$$(i)\; z_{v,v'}^l \le x_v^l, x_{v'}^l, \quad (ii)\; z_{v,v'}^l \ge x_v^l + x_{v'}^l - 1. \tag{13}$$

A similar reformulation is proposed in [45]. The resulting problem can then be solved to optimality using BLP solvers like gurobi [17].

4. Evaluation

We evaluate our tracking approach on a new dataset, which is introduced in Section 4.1. The dataset contains challenging sequences captured with a calibrated camera and body-worn IMUs. In Section 4.2 we provide technical details of our tracking approach and assess its performance in Section 4.3. 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.

4.1. VIMPT Dataset

Current benchmarks for MPT do not contain IMU data. Hence, in order to evaluate our approach we recorded a new dataset denoted VIMPT 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. During soccer recordings, two four-person teams in team dresses (see Figure 5(a)) played soccer in a competitive manner. Consequently, these recordings contain a lot of motion, abrupt changes in direction and occlusions, see Figure 5(b)-(c). Hence, tracking challenges arise from non-linear motion and ambiguous appearance information. Furthermore, the soccer sequences were captured from two different viewpoints and differ in recorded game situations.

In addition to the soccer recordings, the VIMPT dataset contains an outdoor sequence recorded at a pedestrian crosswalk in a public park. Actors walk around in natural apparel and meet regularly for short conversations, see Figure 5(d). This sequence serves as a reference to standard benchmarks such as MOT16[32], since it is comparable in terms of motions and scenery. Throughout all sequences, actors regularly left the field of view and got heavily occluded by other actors.

Camera setup. For all sequences a calibrated camera has been mounted to a tripod at a height of approximately 1.8m. Video was captured in landscape at 30Hz and the camera extrinsics were calibrated to a fixed reference point in the scene.

Detections. We used the person detector Faster R-CNN[35] trained on COCO[30] to generate person detections within all frames of the dataset. For all detections, we compute the corresponding 3D position 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[32], we interpolated ground-truth detections for occluded persons.

IMU setup. Throughout all sequences, eight persons were equipped with an IMU. Each sensor was attached at the person's back at hip height. IMU orientation and acceleration are captured at a framerate of 60Hz and we calibrated the inertial reference coordinate frame to the same reference point as for the camera extrinsics.

Training, validation and test split. We split the VIMPT dataset into three disjoint subsets. The longest soccer sequence is splitted into training and validation parts, while the residual six sequences are used for testing and evaluation.

4.2. Tracker Parameters

The following parameters were empirically chosen, using the training data.

Tracklet generation. We generate reliable tracklets by grouping detections using the method of [34]. In order to avoid error propagation, temporally subsequent detections can only be connected if their intersection over union is above 0.7 and the maximal tracklet length is set to 0.5 seconds.

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 [38]. 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 [39] with $p = 0.3$ between the fully connected layers to avoid overfitting. Finally, the network parameters are learned by minimizing the cost function (5), for given ground-truth detections and corresponding IMU heading vectors of the VIMPT training sequence.

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 0.4 seconds, we establish a short-term edge associated to costs derived from box features. Similarly, we establish long-term edges associated to costs derived from acceleration features between all detections of v and v' if the temporal distance is between 0.4 and 5 seconds.

Feature to cost mapping. In order to transform unary and pairwise features to costs, we use different

(a) (b) (c) (d)

Figure 5: Challenges in the VIMPT dataset. (a) Very similar person appearances. (b) Rapid motions and motion blur. (c) Heavy occlusions. (d) Outdoor scene with frequent occlusions.

strategies. For orientation and box features we learn a logistic regression model[14] that predicts optimal costs based on ground-truth trajectories in the training sequence of the VIMPT dataset. This did not work satisfactory 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 $\min f_{\text{acc}}(\mathcal{H}, \mathcal{H}') > \delta$.

4.3. Evaluation

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

Error metrics. We evaluate tracking performance by assessing assignment coherency in terms of ID metrics. According to [36] 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 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[7]. MOTA comprises three different error metrics, namely number of ID switches, false positives and false negatives. Note that MOTA is calculated based on detection existence only. Instead, IDF1 evaluates false positives (negatives) *and* also verifies that the person ID is correct. Thus we consider IDF1 as the more meaningful metric for the VIMPT task. However, MOTA is a well-known metric for MPT and it enables to put the tracking results into 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 rows of Table 1 and Table 2. 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.

Tracker	IDF1 ↑	IDP ↑	IDR ↑	IDs ↓	MOTA ↑
DeepCC[37]	39.4	40.8	38.2	243	28.7
DeepSORT[46]	45.8	49.6	42.4	193	77.1
FWT[18]	28.1	29.7	26.7	489	71.6
VIT	**91.8**	**93.6**	**90.1**	**44**	**86.1**

Table 1: Tracking accuracy soccer sequences.

Tracker	IDF1 ↑	IDP ↑	IDR ↑	IDs ↓	MOTA ↑
DeepCC[37]	64.6	64.5	64.7	18	78.2
DeepSORT[46]	50.5	53.4	48.0	28	**83.5**
FWT[18]	37.6	39.1	36.3	66	82.4
VIT	**88.5**	**89.5**	**78.5**	**22**	81.8

Table 2: Tracking accuracy outdoor recording.

Comparison to vision-based methods. We apply three different state-of-the-art vision-based trackers to the VIMPT dataset, namely FWT[18], DeepSORT[46] and DeepCC[37]. FWT is the current leader of the MOT17[32] 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 person appearance to establish affinities between detection boxes. In order to better analyze the impact of IMUs, and to be independent of appearance

ambiguities, VIT uses simple spatio-temporal features.

During soccer sequences all players of a team wear identical dresses and hence appearance information is very ambiguous. The tracking results shown in Table 1 validate that this is very challenging for all considered state-of-the-art trackers. Respective IDF1 scores vary between 28.1% 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.

Interestingly, for the park sequence our proposed tracker is on par with the other trackers when MOTA is considered. However, the IDF1 score is still higher indicating that people specific trajectories are recovered more accurately by VIT.

However, the comparison of VIT to vision-based trackers is not completely fair. The number of tracked people for these approaches is not fixed, which is the case for VIT. Consequently, the presented evaluation is more of a qualitative nature and this should be kept in mind when judging absolute numbers.

Influence of IMU features. In order to investigate the influence of orientation and acceleration measurements on the tracking result we report tracking accuracy of three tracker variants: VT, VT+Acc and VT+Ori. We evaluate all trackers on the full VIMPT dataset and show the results in Table 3. VT uses only box features with all costs related to IMU data set to zero. It obtains an IDF1 score of 44.9%, which is approximately 50% worse compared to VIT. VT+Acc extends VT by taking the acceleration feature into account. Tracking accuracy remains almost identical to VT, indicating that simple rejection of very implausible hypothesis pairs is not sufficient in this case. In contrast, incorporating orientation information to VT, denoted as VT+Ori, leads to a significant increase in tracking accuracy yielding an IDF1 score of 88.9%. Hence, orientation consistency in combination with the simple motion model are key to disambiguate tracklet assignments and help to correctly reject most of implausible hypotheses. 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 Visual Heading Network accuracy by computing the relative number of predicted heading vectors $\hat{\mathbf{n}}_d$ 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 4 the network predicts orientation with high accuracy and is able to generalize to unseen images. Since the orientation feature has shown to be very discriminative, the VHN is key to our proposed tracking approach.

Tracker	IDF1 ↑	IDP ↑	IDR ↑	IDs ↓	MOTA ↑
VT	44.9	44.9	44.9	266	65.2
VT+Acc	45.0	44.9	45.1	256	65.0
VT+Ori	88.9	89.9	87.9	79	82.9
VIT	**91.2**	**92.9**	**89.6**	**66**	**85.3**

Table 3: Tracking accuracy of three tracker variants and our proposed tracker (VIT), evaluated on all sequences.

	$\leq 45°$	$\leq 30°$
Train	97.2%	88.8%
Test	96.2%	88.1%

Table 4: Training and test accuracy of the Visual Heading Network. We provide the relative number of heading errors within a threshold of $\epsilon \in \{30°, 45°\}$.

Identification accuracy. According to [36] 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 persons trajectory is assigned to the person label $j \in \mathcal{L}$ defined by the corresponding IMU 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.

5. Conclusions

Combining video and IMU measurements to obtain accurate long-term trajectories is a challenging task. In this work we propose a graph labeling formulation to assign tracklets in the video to corresponding IMU devices. In our experiments, we show that the proposed tracker accurately tracks multiple people even under dynamic motions and heavy occlusions. This demonstrates the potential of the VIMPT setting. As a by-product, we obtain the assignment of each trajectory to the corresponding identity for free. Hence, if a situation at hand allows to equip people with IMUs, our approach represents a useful alternative to appearance-based trackers.

References

[1] A. Alahi, Y. Boursier, L. Jacques, and P. 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.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] A. Alahi, A. Haque, and L. Fei-Fei. Rgb-w: When vision meets wireless. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[5] A. Alahi, V. Ramanathan, K. Goel, A. Robicquet, A. A. Sadeghian, L. Fei-Fei, and S. Savarese. Learning to predict human behavior in crowded scenes. In *Group and Crowd Behavior for Computer Vision*. 2017.

[6] A. Alahi, J. Wilson, L. Fei-Fei, and S. Savarese. Unsupervised camera localization in crowded spaces. In *Proceedings of the IEEE Conference on Robotics and Automation (ICRA)*, 2017.

[7] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *Image and Video Processing*, 2008.

[8] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *Proceedings of the IEEE Conference on Image Processing (ICIP)*, 2016.

[9] M. Camplani, A. Paiement, M. Mirmehdi, D. Damen, S. Hannuna, T. Burghardt, and L. Tao. Multiple human tracking in rgb-depth data: a survey. *IET computer vision*, 11(4):265–285, 2016.

[10] W. Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[11] 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 *Proceedings of the IEEE Conference on Computer Vision (ICCV)*, 2017.

[12] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[14] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[15] M. Fenzi, J. Ostermann, N. Mentzer, G. Payá-Vayá, H. Blume, T. N. Nguyen, and T. Risse. ASEV-automatic situation assessment for event-driven video analysis. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2014.

[16] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza. Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. Georgia Institute of Technology, 2015.

[17] L. Gurobi Optimization. Gurobi optimizer reference manual, 2018.

[18] R. Henschel, L. Leal-Taixé, D. Cremers, and B. 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.

[19] R. Henschel, L. Leal-Taixé, and B. Rosenhahn. Efficient multiple people tracking using minimum cost arborescences. In *German Conference on Pattern Recognition (GCPR)*, 2014.

[20] R. Henschel, L. Leal-Taixé, B. Rosenhahn, and K. Schindler. Tracking with multi-level features. *arXiv preprint arXiv:1607.07304*, 2016.

[21] W. Jiang and Z. Yin. Combining passive visual cameras and active imu sensors to track cooperative people. In *International Conference on Information Fusion (Fusion)*, 2015.

[22] A. R. Jimenez, F. Seco, C. Prieto, and J. 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.

[23] E. S. Jones and S. Soatto. Visual-inertial navigation, mapping and localization: A scalable real-time causal approach. *The International Journal of Robotics Research*, 30(4):407–430, 2011.

[24] M. Keuper, S. Tang, B. Andres, T. Brox, and B. Schiele. Motion segmentation & multiple object tracking by correlation co-clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[25] C. Kim, F. Li, and J. M. Rehg. Multi-object tracking with neural gating using bilinear lstm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[26] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[27] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese. Learning an image-based motion context for multiple people tracking. In *Proceedings of*

the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[28] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *ICCV Workshop on Modeling, Simulation and Visual Analysis of Large Crowds (ICCVW)*, 2011.

[29] E. Levinkov, J. Uhrig, S. Tang, M. Omran, E. Insafutdinov, A. Kirillov, C. Rother, T. Brox, B. Schiele, and B. Andres. Joint graph decomposition & node labeling: Problem, algorithms, applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[31] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy. Learning to track and identify players from broadcast sports videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013.

[32] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, Mar. 2016. arXiv: 1603.00831.

[33] S. Pellegrini, A. Ess, K. Schindler, and L. 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.

[34] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[35] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NIPS)*, 2015.

[36] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop on Benchmarking Multi-Target Tracking (ECCVW)*, 2016.

[37] E. Ristani and C. Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[40] S. Tang, B. Andres, M. Andriluka, and B. Schiele. Multi-person tracking by multicut and deep matching. In *ECCV Workshop on Benchmarking Multi-Target Tracking (ECCVW)*, 2016.

[41] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple people tracking by lifted multicut and person reidentification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[42] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[43] T. von Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and imus. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[44] T. von Marcard, B. Rosenhahn, M. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics)*, 2017.

[45] L. J. Watters. Reduction of integer polynomial programming problems to zero-one linear programming problems. *Operations Research*, 15(6):1171–1174, 1967.

[46] N. Wojke and A. Bewley. Deep cosine metric learning for person re-identification. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[47] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[48] S. Yi, H. Li, and X. Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[49] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[50] J. Zhu, H. Yang, N. Liu, M. Kim, W. Zhang, and M.-H. Yang. Online multi-object tracking with dual matching attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.