Robust Real-Time Tracking of Multiple Objects by Volumetric Mass Densities

Horst Possegger Sabine Sternig Thomas Mauthner Peter M. Roth Horst Bischof
Institute for Computer Graphics and Vision, Graz University of Technology
{possegger,sternig,mauthner,pmroth,bischof}@icg.tugraz.at

Abstract

Combining foreground images from multiple views by projecting them onto a common ground-plane has been recently applied within many multi-object tracking approaches. These planar projections introduce severe artifacts and constrain most approaches to objects moving on a common 2D ground-plane. To overcome these limitations, we introduce the concept of an occupancy volume *– exploiting the full geometry and the objects' center of mass – and develop an efficient algorithm for 3D object tracking. Individual objects are tracked using the local mass density scores within a particle filter based approach, constrained by a Voronoi partitioning between nearby trackers. Our method benefits from the geometric knowledge given by the* occupancy volume *to robustly extract features and train classifiers on-demand, when volumetric information becomes unreliable. We evaluate our approach on several challenging real-world scenarios including the public* APIDIS *dataset. Experimental evaluations demonstrate significant improvements compared to state-of-the-art methods, while achieving real-time performance.*

1. Introduction

Motivated by numerous applications, such as visual surveillance or sports analysis, considerable research has been made in the area of tracking objects from video sequences. For single object tracking, various successful approaches have been proposed, even for robustly handling heavy changes in appearance (*e.g.*, [2]), or geometry (*e.g.*, [11]). In contrast, multi-object tracking (*e.g.*, [5, 9, 19]) is still a challenging problem. As soon as the object density is high and objects are occluding each other, the positions of single instances cannot be determined reliably.

One way to deal with this problem is to take advantage of multiple cameras. In general, these approaches (*e.g.*, [8, 10, 13, 16, 17]) assume overlapping views observing the same 3D scene by exploiting constraints like objects moving on a common ground-plane, a known number of objects, or that two objects cannot occupy the same position at the same time. These constraints are typically referred to as *closed-world assumptions* [14]. Very often, such methods apply change detection in a first step to estimate the foreground likelihood of each pixel (*e.g.*, [10, 16, 17]). Then, this information is fused exploiting the common ground-plane assumption by either computing a score map [10, 16] or by estimating axes intersections [17].

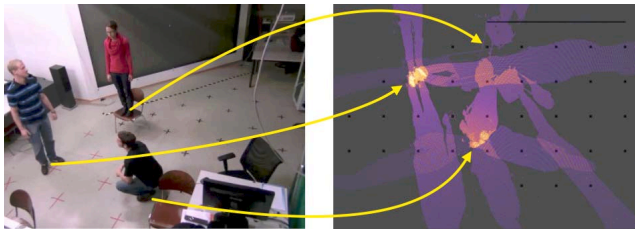
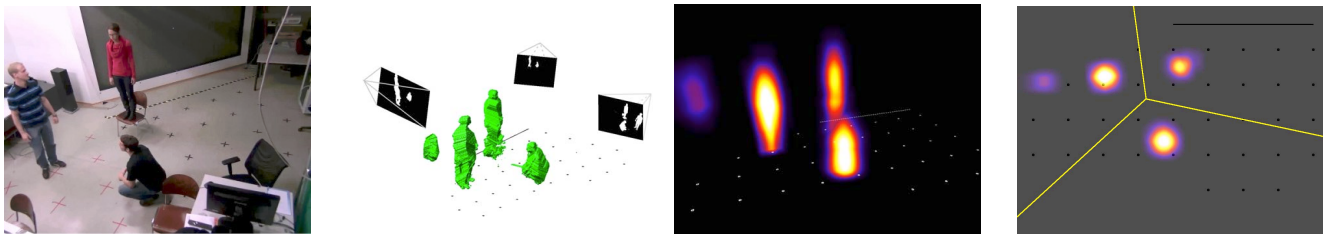


Figure 1: Homography-based approaches (*e.g.*, accumulating projections of foreground segmentations at a common ground-plane) often cause severe artifacts and cannot handle out-of-plane motion (*e.g.*, the person standing on the chair).

One of the main limitations of these methods, as illustrated in Figure 1, is that the planar projections are only valid for the ground-plane, resulting in unreliable projections for points not lying on the ground-plane. Furthermore, these projections generate ghosting artifacts that have to be handled. In order to overcome these limitations, epipolar constraints (*e.g.*, [25]) or volumetric 3D reconstructions (*e.g.*, [6, 12, 21]) are exploited in the tracking process.

Inspired by the idea of using 3D scene structure for multiple object tracking, we propose a robust and real-time capable approach relying on geometric information as a primary cue and using appearance information only on-demand, as opposed to [12, 21]. For that purpose, we introduce the concept of an *occupancy volume*, which is based on local mass densities of a coarse 3D reconstruction of the objects' visual hull. The usage of the local mass density reduces noise and artifacts of the visual hull. This allows to derive an occupancy map, which represents the objects' mass center on the ground-plane for robustly estimating the objects' (x, y) coordinates using a particle filter approach

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.310



(a) Input image. (b) Reconstructed visual hull. (c) Reprojected occupancy volume. (d) Derived occupancy map.

Figure 2: We reconstruct the visual hull from foreground segmentations of input images (a,b), which allows for computing the occupancy volume visualized in (c), where bright colors indicate high local mass densities. The occupancy volume allows for deriving an occupancy map (d) used for robust tracking using particle filtering in combination with Voronoi partitioning.

in combination with Voronoi partitioning. The corresponding z coordinate is then determined using the occupancy volume in a subsequent step. Therefore, in contrast to existing approaches, we are not limited to objects moving on a common ground-plane, which allows for robust tracking of complex scenes, *e.g.*, people stepping on ladders, jumping, etc. Additionally, we exploit the 3D scene structure in combination with the tracking results to on-line collect samples for each individual object. This appearance information can be used to resolve collisions by training discriminative appearance models on-demand.

2. Volumetric Tracking

Tracking algorithms which project 2D image information onto a common ground-plane suffer from ghosting artifacts introduced from the planar projections as shown in Figure 1. To overcome this problem, in the following, we propose a novel multiple camera, multiple object tracking approach exploiting 3D geometric information, which is illustrated in Figure 2.

As a first step, we generate an occupancy volume (see Figure 2c) based on the local mass densities of the 3D visual hull reconstruction (see Figure 2b), which will be introduced and discussed more detailed in Section 2.1. From these occupancy volumes we then estimate occupancy maps (see Figure 2d) and perform the actual tracking step, which is split into two parts to significantly reduce the computational complexity. First, we estimate the targets' coordinates within the Cartesian plane, followed by the estimation of the corresponding z coordinates. Therefore, we refer to this as (2+1)D tracking, which allows for an efficient approximation of an otherwise computationally expensive search within the 3D hypotheses space. By exploiting the 3D occupancy volume, we are able to obtain exact 3D location estimates and furthermore, are not constrained by the common ground-plane assumption. This is described in more detail in Section 2.2. Additionally, as discussed in Section 2.3, we exploit the 3D scene structure to collect samples for learning an appearance model on-demand to resolve ambiguous situations, where a correct assignment solely based on the geometric information cannot be ensured, *e.g.*, whenever targets move too close to each other (see Figure 3).

2.1. 3D Occupancy Volume

Given the foreground segmentations of each camera view (*e.g.*, obtained from standard background subtraction techniques, such as [24]), we reconstruct the visual hull [20]. For that purpose, we adapt Shape from Silhouette [7, 22], to be applicable for reconstructing the visual hulls of objects not visible in all views. Note that Shape from Silhouette is able to handle the constraints imposed by standard multiple camera networks, *i.e.*, wide baselines, significantly different viewing angles (opposing camera positions), and a low number of overlapping views.

In order to reconstruct the visual hull, the scene is discretized into a set $\mathcal{V}$ of $x \times y \times z$ voxels. Every voxel $v_i \in \mathcal{V}$ is reprojected into each camera view where the voxel is visible and set to occupied, *i.e.*, $v_i = 1$, if it projects into the foreground silhouettes, or carved away (*i.e.*, set to background, $v_i = 0$) otherwise. A major advantage of such a voxel-based representation is that it imposes no assumption about the scene planarity, *e.g.*, no common ground-plane is assumed, and thus are perfectly suited for tracking scenarios, where the objects of interest exhibit challenging poses, as can bee seen in Figure 2a.

The visual hull reconstruction is sensitive to noise, *i.e.*, missing or false positive foreground segmentations cause holes in the volume or ghost artifacts. To overcome this problem, we propose an occupancy volume which incorporates information about the voxel's neighborhood. Thus, we are more robust to noisy reconstructions. The 3D occupancy volume can be derived from the visual hull by computing the local mass density m for every voxel v_i, as

$$m(v_i) = \frac{\sum_{v_j \in N_{v_i}} v_j}{|N_{v_i}|}, \tag{1}$$

where the local neighborhood N_{v_i} depends on the objects of interest. For example when considering persons, one can observe that people tend to align their torso upright, *e.g.*, while standing, walking, and even while crouching. Thus, for the task of tracking humans, we define the neighborhood by a cuboid as

$$N_{v_i} = \Big\{ v_j \mid |v_{j,x} - v_{i,x}| \leq r \wedge |v_{j,y} - v_{i,y}| \leq r \wedge |v_{j,z} - v_{i,z}| \leq \frac{h}{2} \Big\}, \quad (2)$$

where $v_{i,x}, v_{i,y}, v_{i,z}$ denote the x, y, z coordinate of the i-th voxel, respectively. By choosing $h \geq 3r$, we increase the emphasis on the vertical neighborhood and thus incorporate the upright alignment of the human torso. Furthermore, by defining the neighborhood relationship as an axis-aligned cuboid, we can use efficient integral image representations for computing the mass densities. The mass density defines a likelihood relationship on the position of an object's center, *i.e.*, the objects' mass centers correspond to high local density values within the occupancy volume.

Although ghost artifacts may occur during reconstruction of the visual hull, their effects are significantly reduced by computing the local mass densities (*e.g.*, see Figure 2c). In general, the mass densities of ghosts vary over time, *i.e.*, these masses are unstable and their masses are lower compared to real objects. The lower mass densities in combination with the closed-world assumption that objects enter and leave the scene at known locations (*i.e.*, they cannot suddenly appear in the middle of the scene) allow for handling ghost artifacts robustly. See Section 2.4 for more details.

2.2. Tracking using the Occupancy Volume

Now, having estimated the occupancy volume, we derive a top view occupancy map $\mathcal{M}$ by assigning the maximum local mass density value along the z axis for a given (x, y) coordinate (see Figure 2d). The actual tracking step is then performed using a particle filtering approach [15] on $\mathcal{M}$. We therefore estimate the target state $\mathbf{x}_t^i = [x, y, v_x, v_y]^\top$ of each object, where (x, y) is the object's location within the Cartesian plane, and (v_x, v_y) describes the object's velocity. Given the mass density observations $\mathbf{z}_t$ of the occupancy map, the posterior probability $p(\mathbf{x}_t^i|\mathbf{z}_t)$ is approximated using a finite set of weighted particles $\{\hat{\mathbf{x}}_t^i, w_t^i\}$.

The particle filter sketched so far works well for single instances, however, collisions of multiple objects cannot be handled. In fact, if objects move close to each other, the respective modes at the occupancy map may coalesce into a single blob, once their visual hulls cannot be separated. Thus, collision handling techniques, such as the iterative repulsion scheme [26], are required. However, by exploiting the assumption that multiple objects cannot occupy the same location in space at the same time, inspired by [18], we can use an efficient approach based on Voronoi partitioning of the hypotheses space (see Figure 2d).

Using the current set of N coordinate estimates $\mathcal{P} = \{P_1, \ldots, P_N\}, P_i = (x_i, y_i)$, we partition the occupancy map $\mathcal{M}$ into a set $\mathcal{C}$ of pairwise-disjoint convex regions $\mathcal{C}_i = \{m \in \mathcal{M} \mid d(m, P_i) \leq d(m, P_j), \forall j \neq i\}$, where $d(\cdot)$ is the Euclidean distance function. According to [18], given the current Voronoi partitioning, the objects' states become conditionally independent and the posterior probability conditioned on $\mathcal{C}$ can be formulated as $p(\mathbf{X}_t|\mathbf{z}_{1:t}, \mathcal{C}_t) = \Pi_i p(\mathbf{x}_t^i|\mathbf{z}_{1:t}, \mathcal{C}_t)$, where $\mathbf{X}_t$ is the concatenation of all objects' states, *i.e.*, the joint-state. This implies that given the Voronoi partitioning, each object can be tracked by a single-object tracker restricted to its corresponding partition. In order to restrict a particle filter's state transition to its respective partition, we use a mask function derived from the partitioning. The partitioning is of special importance if a target is fully occluded by other objects, *i.e.*, not visible in any camera view. Hence, the particle filter keeps the correct position and cannot drift to nearby modes on the occupancy map.

After estimating the objects' locations within the xy plane, the final step to obtain the full 3D coordinates is to compute the corresponding z coordinate. Therefore, we search for the mass center along the z axis within a local neighborhood of the corresponding xy estimate. This additionally allows for correctly tracking objects which exhibit out-of-plane motion.

2.3. Resolving Geometric Ambiguities

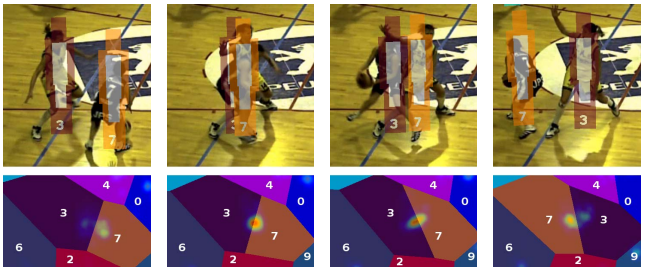
So far, the proposed algorithm operates solely on the geometric information derived from the binary foreground segmentations. However, in real-world scenarios it is often not possible to correctly assign identities using pure geometric information (see Figure 3). To cope with this problem, we develop a *merge-split* approach.

First, we identify potential conflicts between the objects $\mathcal{Q}_i = \{j \mid d(P_i, P_j) < \tau_c, \forall j \neq i\}$, where a robust identity assignment for objects within a radius τ_c on the occupancy map cannot be guaranteed based on the geometric information. Second, for each involved object we train a discriminative one-vs-all classifier, considering only conflicted objects, using on-line collected training samples. We use L2-regularized logistic regression classifiers which solve the unconstrained optimization problem

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C \sum_i \xi\left(\mathbf{w}; \mathbf{f}_i, y_i\right), \quad (3)$$

where $\xi = \log\left(1 + e^{-y_i \mathbf{w}^\top \mathbf{f}_i}\right)$ is the loss function, $\mathbf{f}_i$ are the on-line collected samples for the corresponding objects, $y_i \in \{-1, +1\}$, and $C > 0$ is a penalty parameter.

We exploit the 3D scene structure to identify occlusions in the individual views and thus extract only valuable sam-



(a) $t = 1046$. (b) $t = 1051$. (c) $t = 1054$. (d) $t = 1060$.

Figure 3: Sample merge-split situation at the APIDIS dataset. The colors of the visual results (top) and Voronoi partitions on $\mathcal{M}$ (bottom) correlate. During an attack situation (a,b), the tracker starts drifting (c). By using the discriminative classifier, the trajectories are correctly re-established (d), preventing an identity switch.

ples, *i.e.*, samples where the corresponding object is fully visible. As features we use histograms over the hue and saturation channels of the HSV color space extracted at the torso region, as illustrated by the white overlay in Figure 3. The samples are stored in bags of N_f features per object and updated using a first-in-first-out strategy to account for changing appearance of the objects. Furthermore, we keep separate bags for each camera to ensure robustness in spite of different illumination conditions and to avoid complex color calibration.

While several trackers are in a conflicted state, *i.e.*, volumes are merged, their particle filters share one coalesced maximum. To reduce the complexity of resolving the ambiguities, we exploit the Voronoi partitioning. In particular, we search for separate local maxima on $\mathcal{M}$, restricted by the Voronoi partitions of the involved objects (split). Based on the estimated posterior probability of the logistic regression classifiers we can robustly re-assign the conflicted trackers given the appearance information.

2.4. Automatic Initialization and Cancellation

In order to initialize and cancel trajectories, we define entry regions near the entrances of the scenes, similar to related approaches, such as [3]. This also conforms to the closed-world assumption that objects cannot suddenly appear at the middle of the scene, as applied to reduce the effect of ghosts in the visual hull reconstruction.

For the automatic initialization, we observe the occupancy map at the defined entry areas by extracting maximally stable extremal regions [23]. For each candidate region, we compare whether its mass density corresponds to that of an average human. Before assigning a new tracker, we match the appearance model of the entering person against those of objects which left the scene previously by using the χ^2 distance. Upon a valid match, the entering object is assigned the known identity, otherwise a new trajectory is initialized.

Dataset	N_C	Frames	N_O	FPS	Resolution
APIDIS	7	1500	12	25	1600×1200
CHAP	4	3760	5	20	1024×768
LEAF 1	4	1800	4	20	1024×768
LEAF 2	4	2400	5	20	1024×768
MUCH	4	2400	5	20	1024×768
POSE	4	1820	6	20	1024×768
TABLE	4	1760	5	20	1024×768

Table 1: Dataset characteristics indicating the number of cameras N_C, the total number of frames, the maximum number of simultaneously visible objects N_O, as well as the frame rate (FPS) and the resolution of the video streams.

3. Results and Evaluations

In the following, we demonstrate our proposed multiple object tracker on several challenging real-world people tracking scenarios.

3.1. Datasets

To evaluate our approach and to compare it to the state-of-the-art, we use the publicly available APIDIS[1] dataset and six additional scenarios which we provide[2] for further academic use. The latter were recorded at our laboratory with a tracking region of approximately $7\,\text{m} \times 4\,\text{m}$, using 4 static Axis P1347 cameras. Although the cameras were placed slightly above head-level, at approximately $2.9\,\text{m}$, all sequences exhibit significant occlusions. Table 1 summarizes the technical characteristics of the evaluated datasets, which impose the following challenges:

APIDIS. The public APIDIS dataset shows a basketball game monitored by 7 cameras. This dataset contains various challenges like heavy occlusions, densely crowded situations as well as complex articulations, or abrupt motion changes. Further challenges are caused by the similar appearance of all players of a team, as well as strong shadows and reflections on the floor. Furthermore, some cameras share almost the same viewpoint, *e.g.*, cameras 1 and 7, which provides nearly no additional visual information.

Similar to [1], we evaluate the performance on the left-half of the basketball court, as this side is covered by the larger number of cameras, *i.e.*, cameras 1, 2, 4, 5, and 7. This results in a tracking region of about $15\,\text{m} \times 15\,\text{m}$.

Changing Appearances (CHAP). This sequence depicts a standard surveillance scenario, where 5 people move unconstrained within a laboratory. Throughout the scene, the

[1] http://www.apidis.org/Dataset/
[2] http://lrs.icg.tugraz.at/download#lab6

people change their visual appearance by putting on jackets with significantly different colors than their sweaters. Since people move close to each other after changing their appearance, these situations impose additional challenges to color based object tracking approaches, as fixed color models cannot deal with changing appearances.

Leapfrogs (LEAF 1 & 2). These scenarios depict leapfrog games where players leap over each other's stooped backs. Specific challenges of these sequences are the spatial proximity of players, out-of-plane motion, and difficult poses. Furthermore, two people may share the same xy position while performing a leapfrog which violates the closed-world assumption used for the Voronoi partitioning, as discussed in Section 2.2.

Musical Chairs (MUCH). This sequence shows 4 people playing *musical chairs* (also known as *Going to Jerusalem*) and a non-playing moderator who starts and stops the recorded music. Due to the nature of this game, this sequence exhibits fast motions, as well as crowded situations, *e.g.*, when all players race to the available chairs. Furthermore, sitting on the chairs is a rather unusual pose for typical surveillance scenarios and violates the commonly used constraint of standing persons. Additionally, regarding background modeling, there are dynamic background items, *i.e.*, the chairs which are removed after each round, as well as a static foreground object, *i.e.*, the moderator is standing almost still throughout the whole sequence.

POSE. This sequence shows up to 6 people in various poses, such as standing, walking, kneeling, crouching, crawling, sitting, and stepping on ladders. Additionally to these poses, which again violate common tracking assumptions such as upright standing pedestrians or a common ground-plane, a changing background illumination causes further challenges w.r.t. robust foreground segmentation.

TABLE. This scenario exhibits significant out-of-plane motion as up to 5 people walk and jump over a table. Additional challenges are introduced by densely crowded situations and frequent occlusions.

3.2. Evaluation Metrics

For evaluation, we compute the standard CLEAR multiple object tracking performance metrics [4], *i.e.*, *Multiple Object Tracking Accuracy* and *Precision* (MOTA and MOTP). The precision metric MOTP evaluates the alignment of true positive trajectories w.r.t. the ground truth. We compute the distance between tracker hypotheses and annotated ground truth objects on the ground-plane to allow a comparison between different approaches. The reported MOTP values are measured in meters, where lower values indicate a better alignment with the ground truth. The accuracy metric MOTA is derived from 3 error ratios, namely the ratio of false positives, the ratio of false negatives (*i.e.*, missed objects), as well as the ratio of identity switches. Higher MOTA values indicate a better performance, with 1 representing a perfect tracking result.

We manually annotated every 10th frame for the laboratory scenarios and used the provided ground truth data for the APIDIS dataset. A tracker hypothesis is considered valid if it lies within a radius defined by a distance threshold τ_d of an annotated ground truth position. Note that this distance threshold also defines the upper bound on the reported precision metric MOTP.

3.3. Comparison to State-of-the-Art

We compare our proposed tracking algorithm to the state-of-the-art K-Shortest Paths (KSP) tracker[3] [3]. This tracker operates on a discretized top view representation (grid) and uses peaked probabilistic occupancy maps, which denote the probability that an object is present at a specific grid position. Similar to the original formulation, we obtain the input probability maps using the publicly available implementation[4] of the probabilistic occupancy map (POM) detector [10].

In order to ensure a fair comparison, we use the same foreground segmentations as input to both, our tracking algorithm and the POM detector. For KSP/POM, we divide the top view representation into a grid of $40\,\text{cm} \times 40\,\text{cm}$ cells. The spatial distance between the cell centers was varied from $10\,\text{cm}$ to $20\,\text{cm}$. We set the maximum number of iterations for the POM detector to 1000 and varied its input parameters σ (which accounts for the quality of the foreground segmentations) and the prior probability. Based on the POM results, we additionally evaluated the KSP tracker with varying input parameters, *i.e.*, different limits on the maximum movement between consecutive frames, as well as different entry point setups. The best performing results are reported and used as a baseline for comparison.

3.4. Results and Discussion

Table 2 lists the performance metrics on the individual datasets, while illustrative results[5] are shown in Figure 4. For computing the metrics, we set the distance threshold to $\tau_d = 0.5\,\text{m}$. As can be seen from the overall scores, our proposed algorithm achieves state-of-the-art performance at standard visual surveillance scenarios (*e.g.*, CHAP and LEAF 1), whereas we outperform the KSP tracker at more complex scenarios, *i.e.*, APIDIS, LEAF 2, MUCH, POSE, and TABLE.

[3] http://cvlab.epfl.ch/software/ksp/

[4] http://cvlab.epfl.ch/software/pom/

[5] Additional tracking results are included in the supplemental material.

Dataset	τ_d [m]	Algorithm	MOTP [m]	MOTA	TP	FP	FN	IDS	FPS	N_A
APIDIS	0.50	Proposed	**0.205**	**0.675**	**656**	**88**	**172**	**9**	4.42	27
		Prop. w/o Color	0.211	0.597	625	121	202	10	6.16	-
		KSP/POM	0.231	0.490	607	156	220	46	80.70, 0.03	-
CHAP	0.50	Proposed	**0.102**	**0.994**	1555	**2**	**6**	**1**	9.89	9
		Prop. w/o Color	**0.102**	0.719	1316	193	241	4	12.67	-
		KSP/POM	0.167	0.952	**1607**	50	21	7	43.49, 0.02	-
LEAF 1	0.50	Proposed	**0.107**	**0.991**	464	**2**	2	**0**	9.88	14
		Prop. w/o Color	**0.107**	0.721	436	83	44	7	10.34	-
		KSP/POM	0.169	0.976	**495**	6	**1**	5	63.84, 0.04	-
LEAF 2	0.50	Proposed	**0.097**	**0.916**	**930**	**41**	**41**	**0**	7.65	48
		Prop. w/o Color	0.116	0.727	856	115	117	34	9.04	-
		KSP/POM	0.175	0.819	913	87	66	24	229.77, 0.05	-
MUCH	0.50	Proposed	**0.111**	**0.977**	**783**	**9**	**9**	**0**	12.08	12
		Prop. w/o Color	0.116	0.736	694	99	99	11	13.21	-
		KSP/POM	0.218	0.754	770	139	32	26	185.28, 0.06	-
POSE	0.50	Proposed	**0.123**	**0.944**	**485**	**14**	**14**	**0**	10.27	12
		Prop. w/o Color	0.128	0.822	456	42	44	3	12.99	-
		KSP/POM	0.201	0.555	427	156	31	17	132.49, 0.05	-
TABLE	0.50	Proposed	**0.112**	**0.898**	**599**	**30**	**28**	**6**	8.03	34
		Prop. w/o Color	0.120	0.818	577	56	51	7	9.60	-
		KSP/POM	0.210	0.719	573	105	58	14	208.51, 0.07	-

Table 2: Performance evaluation. For each evaluated dataset, we report the precision metric MOTP (lower is better) and accuracy metric MOTA (higher is better), as well as the total number of true positives (TP), false positives (FP), false negatives (misses, FN), and identity switches (IDS). The best values for each evaluation and each criterion are highlighted. Furthermore, we report the runtime performance in frames per second (FPS), as well as the total number of ambiguous situations N_A, *i.e.*, how often groups of people cannot be distinguished by the geometric information alone (only applicable for the proposed method). Please refer to the text for details.

Similar to [1], we observed a large number of false positives of the POM detector if noisy foreground segmentations are used as input, *e.g.*, caused by changing illumination. Furthermore, in situations where people exhibit challenging poses, missed detections occur frequently. In such situations, the KSP tracker is often not able to link the true positive detections correctly or starts drifting after several frames of missed detections. These issues can be seen by the significantly lower tracking accuracy at the APIDIS, POSE, and TABLE scenarios. In contrast to KSP, our approach is able to handle such complex poses and articulations more robust by exploiting the volumetric information.

Considering the high number of identity switches, the KSP tracker obviously suffers from the missing color information, especially in crowded scenarios. For fair comparison, we evaluated the proposed approach without discriminative appearance models for resolving geometrically ambiguous situations (reported as *Prop. w/o Color*), *i.e.*, trajectory assignment is solely based on the geometric information derived from the occupancy volume. As the local mass densities provide valuable cues for tracking, we still achieve better performances on more complex scenarios compared to the KSP approach, even without using additional color information. By additionally using a discriminative classifier to resolve these ambiguous situations, we achieve excellent tracking results, especially w.r.t. the number of identity switches - *e.g.*, the single identity switch at the CHAP scenario occurs after a person leaves the tracking region, changes his clothes outside, and then re-enters the scene.

Regarding the precision metrics, the proposed approach achieves very accurate results, *i.e.*, the average distance between real object positions and estimated positions is approximately $10\,\text{cm}$. Since the KSP tracker is based on a discretized top view representation, it is constrained by the spatial resolution of the grid. However, as it operates offline on a graph built over all frames, it cannot handle arbitrarily dense grids due to memory limitations. In contrast, our voxel-based approach operates on-line and can be used with high resolution volumes, *i.e.*, we set the voxel size to $1\,\text{cm} \times 1\,\text{cm} \times 5\,\text{cm}$ for the evaluated scenarios.

As can be seen from the reported metrics on the APIDIS dataset, we still achieve very accurate and precise tracking results, despite the challenges caused by shadowing effects and heavy reflections, as well as the complex and fast movement of the players. Although the on-line sample collection facilitates correctly tracking players of different teams, identity switches occur due to the similar appearance of players within a team. Therefore, more discriminative vi-

sual features specifically designed for sports analysis, *e.g.*, obtained via jersey number recognition, may further improve the overall performance on this scenario.

3.5. Runtime Performance

Table 2 contains the runtime performance in fps evaluated on a standard PC with a 3.2 GHz Intel CPU, 16 GB RAM, and a GeForce GTX580. We achieve frame rates of up to 12 fps for standard tracking scenarios, although only the visual hull reconstruction and the occupancy volume are computed on the GPU, exploiting the inherent parallelism. Note that the reconstruction volume used for the APIDIS dataset is approximately 8 times larger than for the remaining scenarios, which causes the lower frame rate.

The KSP tracker achieves very high frame rates due to the efficient shortest path computation. We report the runtimes for those KSP/POM configurations which achieve the best tracking performance. Thus, the reported frame rates vary for scenarios with similar input data, as the KSP runtime depends on the spatial grid density.

In contrast, the POM detector exhibits a significantly lower frame rate caused by the high resolution of the input images, as well as the required parameter configurations to handle the noisy foreground segmentations. Hence, the combination of KSP and POM does not achieve real-time capability on the evaluated scenarios. Furthermore, the KSP tracker requires the detection probabilities for all time steps in advance for constructing the underlying graph structure. Thus, it can only be computed off-line, whereas the proposed approach works completely on-line at high frame rates.

4. Conclusion

We proposed a real-time capable multi-object tracking approach based on local mass densities of visual hull reconstructions. In contrast to existing tracking approaches for calibrated camera networks with partially overlapping views, we are not constrained by the common ground-plane assumption and additionally reduce artifacts rising from noisy foreground masks. In particular, individual objects are tracked using the local mass density scores within a particle filter framework, constraining nearby trackers by a Voronoi partitioning. Furthermore, we continuously exploit the reconstructed 3D information to robustly extract features on-line. These features are used to train discriminative classifiers in situations where pure geometric information becomes unreliable. To demonstrate the benefits of our proposed approach, we generated several challenging datasets and additionally evaluated our approach on the publicly available APIDIS basketball dataset. In both cases, state-of-the-art methods can be outperformed in terms of precision and accuracy, as well as runtime. Future work will concentrate on extracting different features to allow for more robust handling of objects with similar appearance (*e.g.*, relevant for APIDIS).

Acknowledgments

This work was supported by the Austrian Science Foundation (FWF) under the projects Advanced Learning for Tracking and Detection in Medical Workflow Analysis (I535-N23) and MASA (P22299).

References

[1] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst. Sparsity Driven People Localization with a Heterogeneous Network of Cameras. *JMIV*, 41(1-2):39–58, 2011.

[2] B. Babenko, M.-H. Yang, and S. Belongie. Robust Object Tracking with Online Multiple Instance Learning. *PAMI*, 33(7):1324–1338, 2011.

[3] J. Berclaz, F. Fleuret, E. Türetken, and P. Fua. Multiple Object Tracking using K-Shortest Paths Optimization. *PAMI*, 33(9):1806–1819, 2011.

[4] K. Bernardin and R. Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP JIVP*, 2008.

[5] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online Multi-Person Tracking-by-Detection from a Single, Uncalibrated Camera. *PAMI*, 33(9):1820–1833, 2010.

[6] C. Canton-Ferrer, J. R. Casas, M. Pardàs, and E. Monte. Multi-camera multi-object voxel-based Monte Carlo 3D tracking strategies. *EURASIP JASP*, 2011(114), 2011.

[7] K.-M. Cheung, S. Baker, and T. Kanade. Shape-From-Silhouette Across Time Part I: Theory and Algorithms. *IJCV*, 62(3):221–247, 2005.

[8] R. Eshel and Y. Moses. Tracking in a Dense Crowd Using Multiple Cameras. *IJCV*, 88(1):129–143, 2010.

[9] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. Robust Multi-Person Tracking from a Mobile Platform. *PAMI*, 31(10):1831–1846, 2009.

[10] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multi-Camera People Tracking with a Probabilistic Occupancy Map. *PAMI*, 30(2):267–282, 2008.

[11] M. Godec, P. M. Roth, and H. Bischof. Hough-based Tracking of Non-Rigid Objects. In *Proc. ICCV*, 2011.

[12] L. Guan, J.-S. Franco, and M. Pollefeys. Multi-Object Shape Estimation and Tracking from Silhouette Cues. In *Proc. CVPR*, 2008.

[13] M. Hu, J. Lou, W. Hu, and T. Tan. Multicamera Correspondence Based on Principal Axis of Human Body. In *Proc. ICIP*, 2004.

[14] S. S. Intille and A. F. Bobick. Visual Tracking Using Closed-Worlds. In *Proc. ICCV*, 1995.

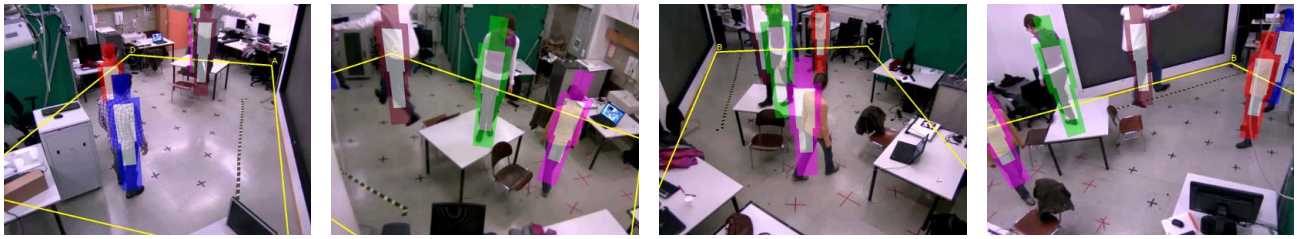
[15] M. Isard and A. Blake. CONDENSATION - Conditional Density Propagation for Visual Tracking. *IJCV*, 29(1):5–28, 1998.

[16] S. M. Khan and M. Shah. Tracking Multiple Occluding People by Localizing on Multiple Scene Planes. *PAMI*, 31(3):505–519, 2009.

(a) APIDIS scenario.

(b) LEAF 2 scenario.

(c) MUCH scenario.



(d) TABLE scenario.

Figure 4: Illustrative tracking results. These situations show the basketball court (a), the leapfrog exercises (b), the musical chairs game (c), and violations of the common ground-plane assumption (d). The yellow rectangle limits the tracking region. Best viewed in color.

[17] K. Kim and L. S. Davis. Multi-camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering. In *Proc. ECCV*, 2006.

[18] M. Kristan, J. Perš, M. Perše, and S. Kovačič. Closed-world tracking of multiple interacting targets for indoor-sports applications. *CVIU*, 113(5):598–611, 2009.

[19] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? *Proc. CVPR*, 2011.

[20] A. Laurentini. The Visual Hull Concept for Silhouette-Based Image Understanding. *PAMI*, 16(2):150–162, 1994.

[21] M. Liem and D. M. Gavrila. Multi-person tracking with overlapping cameras in complex, dynamic environments. In *Proc. BMVC*, 2009.

[22] W. N. Martin and J. K. Aggarwal. Volumetric Descriptions of Objects from Multiple Views. *PAMI*, 5(2):150–158, 1983.

[23] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *Proc. BMVC*, 2002.

[24] N. J. B. McFarlane and C. P. Schofield. Segmentation and tracking of piglets in images. *MVA*, 8(3):187–193, 1995.

[25] A. Mittal and L. S. Davis. M_2Tracker: A Multi-View Approach to Segmenting and Tracking People in a Cluttered Scene. *IJCV*, 51(3):189–203, 2003.

[26] W. Qu, D. Schonfeld, and M. Mohamed. Real-Time Interactively Distributed Multi-Object Tracking Using a Magnetic-Inertia Potential Model. In *Proc. ICCV*, 2005.