Optimizing Long-Term Player Tracking and Identification in NAO Robot Soccer by fusing Game-state and External Video

Anonymous Author(s)
Affiliation
Address
email

Abstract: Monitoring a fleet of robots requires stable long-term tracking with re-identification, which is yet an unsolved challenge in many scenarios. One application of this is the analysis of autonomous robotic soccer games at RoboCup. Tracking in these games requires handling of identically looking players, strong occlusions, and non-professional video recordings, but also offers state information estimated by the robots. In order to make effective use of the information coming from the robot sensors, we propose a robust tracking and identification pipeline. It fuses external non-calibrated camera data with the robots' internal states using quadratic optimization for tracklet matching. The approach in this work is validated using game recordings from previous RoboCup World Cups.

1 Introduction

Robust tracking with stable object identification is a crucial step towards extracting game statistics and improving gameplay in many team sports. While this is usually approached using an external camera only, our application in understanding soccer games played by humanoid robots allows us to fuse this information with measurements from robot-mounted sensors. In this work, we focus on the RoboCup Standard Platform League (SPL) where humanoid NAO robots compete fully autonomously. Game analytics in this setting can offer objective feedback on the algorithms' performance to the teams and help to improve the gameplay.

Our problem differs in multiple ways from the well-known tracking and identification problem in game analytics: RoboCup games are recorded with non-professional uncalibrated camera equipment, robots look identical except for their jerseys, jersey numbers are too small to be detected reliably, and human referees often occlude the camera. To handle these challenges, we propose a long-term tracking pipeline consisting of the following modules:

1. Camera calibration, to estimate camera distortion, intrinsics, and pose relative to the field.
2. Short-term object tracking based on Tracktor [1] and trained on our data to generate tracklets.
3. Optimization-based long-term tracking and player identification by fusing cues from an external camera and the robot sensor data.

2 Related Work

Multi Object Tracking (MOT) describes the tracking of all objects belonging to a given set of categories [2]. In joint tracking and detection approaches, the object detector is a fundamental part of the tracking pipeline [1, 3, 4]. We use Tracktor [1], which follows this paradigm as a building block in our pipeline. Another category of trackers uses detections provided by a separate object detector, followed by solving a data association problem. This framework includes fully deep learning based methods [5, 6, 7] as well as optimization based approaches [8, 9, 10].

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

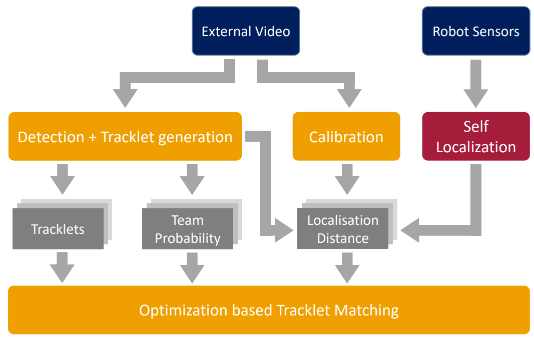

Figure 1: Overview of the proposed approach.

Game Analytics covers the tracking and identification of the players in videos [11] as a key challenge, where MOT is an important component. This can further be aided by robust team detection [11, 12]. A key aspect of our work, usually not addressed in player tracking and identification is the integration of player-mounted sensor data, as it is not easily applicable to human players.

Camera Calibration Tracking and identity assignment requires accurate camera intrinsics and extrinsics. Standard calibration processes generally use point correspondances, robot's motion or calibration patterns to provide accurate intrinsics [13, 14, 15], which cannot be used in our case due to poor point correspondences. Therefore, our approach utilizes a technique proposed by Alvarez et al. [16], which minimizes an energy objective based on rectifying lines present on the field.

3 Method

In this section, we detail our pipeline for consistent player tracking and identification/ Figure 1 provides an overview of the key components and information flow. First, the camera is calibrated to compute its intrinsic and extrinsic parameters with respect to the known soccer field, which is required only once for each video sequence. Then, the tracklets, the team color for each player and all relevant robot state information is extracted. Subsequently, each tracklet is associated with a specific robot player by optimizing a binary quadratic program as described in Section 3.5.

Data and Application: We consider RoboCup SPL matches between humanoid NAO robots, using a dataset comprised of 8 annotated 5000-frame sequences recorded by wide-angle cameras at 30 FPS. The videos were recorded at RoboCup 2019 and 2022, together with the corresponding team communication and game controller logs. Annotations include the bounding box, jersey color and number for each active player and frame. Object detection and classification models and the optimizer hyperparameters are trained on five sequences, and evaluated on the remaining three.

3.1 Camera Calibration and Extrinsics

We assume a static camera over the sequence and compute the median of each pixel over all frames to remove moving objects and obtain a clear view of the field. Then the wide-angle image is undistorted, by estimating the distortion using [16] on detections of field line candidates. We apply the SOLD2 [17] line detector on the undistorted image and obtain intersection points on the field which can then be matched to known field 3D coordinated. The camera pose is computed using P3P [18].

3.2 Multi Object Tracker

To generate tracklets, we use Tracktor [1] with Faster-RCNN [19] and a ResNet-50 backbone. We initialize the model with MS-COCO [20] pretrained weights and fine-tune it on the 5 training sequences of our dataset. The tracklets are robust in easy tracking scenarios without occlusions, but do not cover a whole video. For further processing, each tracklet is projected to field coordinates

Figure 2: Visualization of robots identified by the tracker. The tracking result is represented by bounding boxes and IDs at their top. Ground truth positions are represented by green crosses and corresponding green IDs.

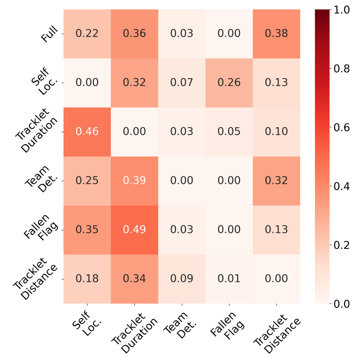

Figure 3: Ablation weigths.

using the estimated camera pose. Subsequently, each projected tracklet is smoothed using a Kalman filter with a constant velocity model.

3.3 Jersey color detection

In the SPL, 9 distinct jersey colors are used. The team colors for a match provide a strong signal to associate tracklets to players from either team. We detect colors using a VGG16 network that assigns a score for each of the team colors used by the two teams in a game.

3.4 Robot States

We furthermore use the robots' states for our matching problem. These include information from the robot sensors as well as the game state:
– Self Localization: The robots calculate their relative position on the field based on their views.
– Fallen Robots: The robots use the IMU sensor information to determine if they have fallen.
– Penalized Robots: The robots removed from the field is a constraint in the optimization problem.

3.5 Global optimization

Occlusions and distractors cause Tracktor to split the idealy long tracks into a large number of shorter tracklets. Therefore, we frame the long-term tracking problem as an assignment of tracklets to a fixed number of player tracks similar to [10] as a constrained quadratic binary optimization problem. We denote the index set of player tracks $I = \{1, ..., N\}$ (with $N = 10$) and generated tracklets $J = \{1, ..., M\}$. The objective is to minimize:

$$H(x) = \sum_{i \in I} \sum_{j \in J} x_{i,j} (O_u + \sum_{l \in L} w^l c^l_{i,j}) + \sum_{i \in I} \sum_{j \in J} \sum_{k \in J} x_{i,j} x_{i,k} (\sum_{p \in P} w^p c^p_{j,k}) \tag{1}$$

where $x_{i,j} \in \{0, 1\}$ are binary optimization variables, with $x_{i,j} = 1$ meaning tracklet j is assigned to track i, L and P the sets of unary and pairwise cost functions with costs $c^l_{i,j}$ and $c^p_{i,j,k}$. w^l and w^p are used to weight different cost terms. The offsets are negative to penalize the trivial solution of assigning nothing ($x_{i,j} = 0 \, \forall i, j$). Two constraints are imposed to ensure feasible matchings:

1. One tracklet can only be assigned to one track.
2. Temporally overlapping tracklets cannot be merged to the same track.

3.6 Cost terms

Different cost terms control the assignment of tracklets to tracks. We use the following terms.

- **Duration**: Penalizes short tracklets, as these are often spurious tracklets.
- **Self-localization**: The distance between the position estimated by a robot from its camera and the position estimated for a tracklet from the external camera.

Full	Self Loc.	Tracklet Duration	Team Det.	Fallen Flag	Penalized Flag	Tracklet Distance
88.1	15.4	51.1	76.5	86.2	76.3	83.3

Table 1: Tracking performance and ablation study. Results are provided in percent MPIR.

Time Frame	30	150	300	900	1800	3600	5400
MPIR	42.5	42.3	40.1	43.8	38.4	38.5	38.5

Table 2: DeepSort Performance with different re-identification time.

- **Jersey color detection**: The score for the color-based team detection throughout a tracklet.
- **Global trajectory continuity**: The pairwise spatial distance between the end of a tracklet and the start of a new temporally close tracklet.

3.7 Reference Method: DeepSORT

To fulfill the task of long-term tracking and player identification under strong occlusions we augment the DeepSORT tracker [21, 22] by a greedy tracklet matching algorithm that matches any new tracklet to the spatially closest inactive track. Constraints are applied to prevent temporarily overlapping tracklets from being assigned to the same track. To provide a strong baseline, we provide this approach with oracle information; the total number of robots that are present in a sequence, which defines the maximum number of tracks as well as the ground truth ID of the first tracklet for each robot. Finally, the best re-identification time for DeepSORT is selected on the testset.

4 Results

We evaluate our approach over a test set containing 3 video sequences of 3 minutes. Each video covers a different game, thus testing our approach under different conditions.

Table 1 shows the Mean Player Identification Recall (MPIR), the ratio of times each player has been identified correctly. The first column shows our full approach. Subsequent columns show ablations, each feature removed separately with the cost term optimized using PSO [23] for each scenario.

With all features we achieve 88.1% MPIR. Removing the robot self localization has the strongest impact with 15.4% MPIR, while removing the fallen robot flag results in the least performance drop. The self-localization is an important feature since it provides information about the position of the robot. The fallen robot flag is unreliable, as it relies on the robot's IMU and a heuristic to detect whether the robot has fallen in the video.

Table 2 shows the performance of our DeepSORT baseline over different reidentification times. The best performance as achieved with 900 frames which corresponds to 30s of video. In this case, the performance is 43.8% MPIR, compared to 88.1% MPIR with our method using all available features.

We further analyze each feature's importance through its weighting, where a higher weight indicates a more important feature. Figure 3 shows the importance of the features in the different ablations. Strong weights are assigned to the self-localization and tracklet duration. Removing these features shows that weighting is redistributed: the noisy fallen robot events are incorporated when no self-localization information is available, as it can provide unique information about a robot's ID.

5 Conclusion

In this work, we presented a sensor fusion based method for tracking multiple similar humanoid robots. We utilize information from both visual data and their own sensors by combining tracklets using a quadratic optimization technique. The method allows automated tracking of robot players over a long time on a stationary video sequence. Open points that we will investigate in the future include the evaluation in more complex environments, the interpolation of tracks during occlusions as well as the extraction of high-level game statistics.

References

[1] P. Bergmann, T. Meinhardt, and L. Leal-Taixe. Tracking Without Bells and Whistles. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 941–951, Seoul, Korea (South), Oct. 2019. IEEE. ISBN 978-1-72814-803-8. doi:10.1109/ICCV.2019.00103.

[2] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. Reid, S. Roth, K. Schindler, and L. Leal-Taixé. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003 [cs]*, Mar. 2020.

[3] S. Karthik, A. Prabhu, and V. Gandhi. Simple Unsupervised Multi-Object Tracking. *arXiv:2006.02609 [cs]*, June 2020.

[4] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz. Joint discriminative and generative learning for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[5] G. Braso and L. Leal-Taixe. Learning a Neural Solver for Multiple Object Tracking. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6246–6256, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi:10.1109/CVPR42600.2020.00628.

[6] X. Zhou, V. Koltun, and P. Krähenbühl. Tracking Objects as Points. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 474–490, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58548-8. doi:10.1007/978-3-030-58548-8_28.

[7] J.-N. Zaech, D. Dai, A. Liniger, M. Danelljan, and L. V. Gool. Learnable online graph representations for 3d multi-object tracking. In *IEEE International Conference on Robotics and Automation, ICRA*, 2022. URL `https://arxiv.org/abs/2104.11747`.

[8] A. Hornakova, R. Henschel, B. Rosenhahn, and P. Swoboda. Lifted Disjoint Paths with Application in Multiple Object Tracking. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4364–4375. PMLR, Nov. 2020.

[9] S. Tang, M. Andriluka, B. Andres, and B. Schiele. Multiple People Tracking by Lifted Multicut and Person Re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3701–3710, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi:10.1109/CVPR.2017.394.

[10] J.-N. Zaech, A. Liniger, M. Danelljan, D. Dai, and L. Van Gool. Adiabatic quantum computing for multi object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8811–8822, June 2022.

[11] T. Yamamoto, H. Kataoka, M. Hayashi, Y. Aoki, K. Oshima, and M. Tanabiki. Multiple players tracking and identification using group detection and player number recognition in sports video. In *IECON 2013-39th Annual Conference of the IEEE Industrial Electronics Society*, pages 2442–2446. IEEE, 2013.

[12] S. Gerke, K. Muller, and R. Schafer. Soccer jersey number recognition using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 17–24, 2015.

[13] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[14] M. Pollefeys, L. Van Gool, and A. Oosterlinck. The modulus constraint: a new constraint self-calibration. In *Proceedings of 13th International Conference on Pattern Recognition*, volume 1, pages 349–353. IEEE, 1996.

[15] T. Probst, K.-K. Maninis, A. Chhatkuli, M. Ourak, E. Vander Poorten, and L. Van Gool. Automatic tool landmark detection for stereo vision in robot-assisted retinal surgery. *IEEE Robotics and Automation Letters*, 3(1):612–619, 2017.

[16] L. Alvarez, L. Gómez, and J. R. Sendra. An algebraic approach to lens distortion by line rectification. *Journal of Mathematical Imaging and Vision*, 35(1):36–50, 2009.

[17] R. Pautrat, J.-T. Lin, V. Larsson, M. R. Oswald, and M. Pollefeys. Sold2: Self-supervised occlusion-aware line description and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11368–11378, June 2021.

[18] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR 2011*, pages 2969–2976. IEEE, 2011.

[19] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, June 2017. ISSN 0162-8828, 2160-9292. doi:10.1109/TPAMI.2016.2577031.

[20] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common Objects in Context. *arXiv:1405.0312 [cs]*, May 2014.

[21] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017. doi:10.1109/ICIP.2017.8296962.

[22] N. Wojke and A. Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756. IEEE, 2018. doi:10.1109/WACV.2018.00087.

[23] R. Poli, J. Kennedy, and T. M. Blackwell. Particle swarm optimization. *Swarm Intelligence*, 1: 33–57, 1995.