# Efficient tracking of team sport players with few game-specific annotations

Adrien Maglo, Astrid Orcesi, Quoc-Cuong Pham

# Efficient tracking of team sport players with few game-specific annotations

Adrien Maglo Astrid Orcesi Quoc-Cuong Pham
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{adrien.maglo, astrid.orcesi, quoc-cuong.pham}@cea.fr

## Abstract

*One of the requirements for team sports analysis is to track and recognize players. Many tracking and re-identification methods have been proposed in the context of video surveillance. They show very convincing results when tested on public datasets such as the MOT challenge. However, the performance of these methods are not as satisfactory when applied to player tracking. Indeed, in addition to moving very quickly and often being occluded, the players wear the same jersey, which makes the task of re-identification very complex. Some recent tracking methods have been developed more specifically for the team sport context. Due to the lack of public data, these methods use private datasets that make impossible a comparison with them. In this paper, we propose a new generic method to track team sport players during a full game thanks to few human annotations collected via a semi-interactive system. Non-ambiguous tracklets and their appearance features are automatically generated with a detection and a re-identification network both pre-trained on public datasets. Then an incremental learning mechanism trains a Transformer to classify identities using few game-specific human annotations. Finally, tracklets are linked by an association algorithm. We demonstrate the efficiency of our approach on a challenging rugby sevens dataset. To overcome the lack of public sports tracking dataset, we publicly release this dataset at https://kalisteo.cea.fr/index.php/free-resources/. We also show that our method is able to track rugby sevens players during a full match, if they are observable at a minimal resolution, with the annotation of only 6 few seconds length tracklets per player.*

## 1. Introduction

Player tracking in team sports consists in detecting and identifying the players in video sequences. It is a necessary task to automate the generation of individual statistics such as ball possession, field position or involvement in play sequences. Player tracking in team sports such as rugby is however a challenging task. Rugby is a sport of physical contact where player occlusions are very frequent on camera during rucks, tackles and scrums. The players can also adopt a wide range of body postures from sprinting to laying on the ground in a foetal position. Players from the same team share a very similar appearance since they wear the same jerseys. Moreover, the number of pixels in which the players are visible is often limited in the case of a TV stream (sometimes with a height below 150 pixels). This prevents the access to fine identification details.


Figure 1. Tracking French players (blue jerseys) in our rugby sevens dataset: a. France / Kenya extract, b. Argentina / France extract, c. France / Chile extract.

Player tracking is a specific Multi-Object Tracking (MOT) problem. MOT has been widely studied in the literature. Security applications have lead to the development of many people tracking approaches. Offline methods use all the frames of the input video sequence to optimize the generation of tracks while online methods target real-time applications by relying only on the current and previous frames to generate the tracks. The most recent frameworks achieve the best performance using deep neural network ar-

chitectures. The availability of large public datasets and challenges such as the MOT challenge [38] allows to fairly train and compare the various approaches.

Some recent people tracking methods have also been proposed for the specific context of team sports: soccer [24, 54], basketball [34] and hockey [47]. These methods often use private datasets specific to their studied sport to get competitive results evaluated on short video clips extracted from a match. Game-specific annotations are required to train a player tracking and identification system to adapt to the player identities and the context of a game. The number of such annotations is an important factor that will determine the success of using such a system in real world scenario. Little focus was made in previous work on the practicability of this annotation process. Consequently, we propose an incremental learning approach to identify players with very few game-specific annotations. Our method is offline: it tracks and identifies players once the game has been completed. It benefits from the closed gallery re-identification (re-ID) hypothesis as, contrary to video surveillance, the number of players is known and limited. Since our method does not use any sport-specific knowledge, it can be applied to any team sport.

Our annotation process consists in several steps. Bounding boxes around all the persons in the frames are first extracted from the input video to generate non-ambiguous tracklets. A tracklet is the uninterrupted sequence of the bounding box images of a single player. Tracklets can have a variable length since a player can enter or leave the camera field of view or be occluded by an other player. At this stage, the user provides few annotations per player to train the tracklet re-ID network. Finally, the obtained tracklet classification scores or appearance features feed an algorithm that look for an optimal association between tracklets and identities.

The contributions of this paper are the following:

We tackle the sparsity of training data in team sport contexts by leveraging generic detection and re-ID datasets. The detection network is only learned on a public dataset. The re-ID network is pretrained on a video surveillance public dataset.

We propose a new architecture based on a Transformer network [46] to classify and generate tracklet appearance features. An incremental learning mechanism using few user interactions trains this model and strengthen the re-ID performances throughout the whole annotation process.

Some datasets have been proposed for basketball [8] and soccer [11] player tracking with multiple static cameras. However, although Deliege et al. [9] are extending their SoccerNet dataset to tracking and re-ID, no dataset with a moving point of view has been made available. We publicly release our rugby sevens tracking dataset composed of single-view videos that can pan, tilt or zoom to follow the action. It is one of the most challenging team sport for tracking on which no approach was tested.

We demonstrate the efficiency of our approach on our dataset. On a full game, it can achieve up to 67.9% detection and identity classification recall when the players are sufficiently visible with only 6 annotations per player.

The paper is organized as follows: Section 2 introduces Related Work. Our method is described in Section 3. Finally, Section 4 provides our results on our challenging rugby sevens dataset, compares them to state-of-the-art methods and analyzes them in an ablative study.

## 2. Related Work

### 2.1. Multiple people tracking

Two categories of MOT algorithms can be distinguished.

**Offline methods** leverage the full sequence of images to globally optimize the generated tracks with a graph paradigm. The vertices are the detections on each frame and the edges are the connections between detections that form tracks. Thus, Zhang et al. [53] uses a minimum cost flow iterative algorithm that models long term occlusions. The approach described by Berclaz et al. [2] takes only an occupancy map of detection as input and applies a k-shortest path algorithm on the flows. More recently, Brasó and Leal-Taix [5] proposed a fully differentiable network that learns both the appearance and geometrical feature extraction as well as the detection association to generate tracks. Hornakova et al. [23] use lifted edges to model long term interactions and generate the optimized solution with a linear programming relaxation.

**Online methods** only use the current and past frames to associate new detections with tracks. They have raised more interest in the literature as they fit real time scenario. Thus, the SORT algorithm [4] uses a Faster R-CNN [40] person detector. Then, a Kalman filter [25] predicts the future positions of each track. The Intersection-Over-Union (IOU) between these predictions and the detected bounding boxes are used as inputs of an Hungarian algorithm that matches the detection with the tracks. ByteTrack [55] achieves state of the art tracking performance with a two steps association algorithm: the first step focuses on high confidence detections while the second step deals with the low confidence ones. The Deep SORT algorithm [49] adds a re-ID network to extract the visual appearance of each person. The input data of the Hungarian algorithm becomes a combination of a Manaholis distance as the spatial term and a cosine distance between the re-ID vectors as the appearance term.

Using distinct networks for detection and re-ID has the advantage of separating the two tasks that may have opposite objectives. The detection task aims at learning common features to recognize humans while the re-ID task aims at learning distinctive features of each individual. However

this may cause scalability issues as each detected bounding box must be independently processed by the re-ID network. Single-shot methods were therefore proposed to generate the bounding boxes coordinates and re-ID vectors with a single network. Thus, Track-RCNN [48] uses a common backbone with specific heads for each task. FairMOT [56] achieves better tracking performance by focusing only on the detection and re-ID tasks. Meinhard et al. [37] use a Transformer architecture.

Applying traditional MOT to team sport players usually leads to many ID switches. Each time a player leaves the vision field or is occluded too much time, a new identity is generated at reappearance. This prevents the reliable generation of individual statistics (see section 4.2.3).

## 2.2. Multiple team sport player tracking and re-identification

### 2.2.1 Tracking

Some tracking methods have been proposed for the context of team sports. For soccer, many approaches performed tracking by first extracting the field regions [1,12,26,33,36, 50]. In the method of Liu et al. [33], an unsupervised clustering algorithm classifies the players among four classes (two teams, referee or outlier). The tracking is formulated as a Markov chain Monte Carlo data association. D'Orazio et al. [12] classify each player with an unsupervised clustering algorithm. The tracking takes as input geometrical and motion information. It is based on a set of logical rules with a merge split strategy. In Xing et al. [50], the observation model of each player is composed of the color histogram in the cloth regions, the size and the motion. The tracking is formulated as particle filtering. Theagarajan and Bhanu [45] used a YOLOv2 [39] network detector and a DeepSORT tracker [49] to identify the player controlling the ball.

All the previous approaches do not build individual appearance signatures per player identities. If a player leaves the camera field of view and re-enter later, he/she will be considered as a new person. This prevents the generation of individual statistics.

### 2.2.2 Re-identification

Jersey number recognition has been studied in the literature to identify team sport players. Ye et al. [52] developed a method based on Zernike moments features [27]. Gerke et al. [15] were the first to use a convolutional neural network to classify jersey numbers from bounding box images of players. It was later combined to spatial constellation features to identify soccer players [14]. To ease the recognition of distorted jersey numbers, Li et al. [30] trained a branch of their network to correct the jersey number deformation before the classification. Liu and Bhanu [32] enabled jersey number recognition only in the relevant zones by detecting body keypoints. For hockey player identification, Chan et al. [7] used a ResNet + LSTM network [19, 22] on tracklet images to extract jersey numbers.

When a single view is available, as in our rugby sevens dataset, jersey numbers are often not visible, partially visible or distorted. Besides, to our knowledge, there is no publicly available training dataset for team sport jersey number recognition. A solution can therefore be to use appearances to re-identify players. Teket and Yetik [44] proposed a framework to identify the player responsible for a basketball shot. Their re-ID network, based on MobileNetV2 [42], is trained with a triplet loss formulation. The framework described by Senocak et al. [43] combines part-based features and multiscale global features to generate basketball player signatures. Both approaches are based, as ours, on the hypothesis of a closed gallery however they use a private dataset to train their model which makes comparisons impossible.

### 2.2.3 Tracking with re-identification

Several methods tracks players by using re-ID features [24,34,47,51,54]. Lu et al. [34] use DPM [13] to detect basketball players. Local features and RGB color histograms are extracted on players for the re-ID. Zhang et al. [54] proposed a multi-camera tracker that locates basketball players on a grid based on a K-shortest paths algorithm [2]. Players are detected and segmented with a network based on Mask R-CNN [17]. Re-ID features are computed thanks to the team classification, jersey number recognition and a pose-guided feature embedding. To track soccer players, Yang et al. [51] iteratively reduced the location and identification errors generated by the previous approach by creating a bayesian model that is optimized to best fit input pixel level segmentation and identification. Hurault et al. [24] use a single network with a Faster R-CNN backbone [40] to detect small soccer players and extract re-ID features. Kong et al. [28] mix player appearance, posture and motion criteria to match new detections with existing tracks. Vats et al. [47] use a Faster R-CNN network [40] to detect hockey players and a batch method for tracking [5]. Specific ResNet-18 networks [19] are used to identify the player teams and jersey numbers.

Most of the approaches presented here [34, 47, 51, 54] train their re-ID or jersey number recognition model with a private dataset.

### 2.2.4 Minimizing the number of annotations

To our knowledge, few previous work focus on the minimization of the game-specific training annotations for re-ID. For example, Lu et al. [34] used a mere 200 labels

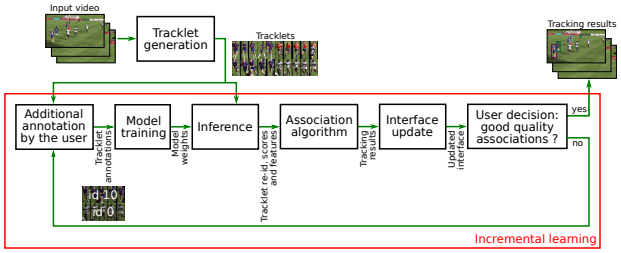


Figure 2. Incremental learning of tracklet classification. The user provides annotations to train the model to correctly classify the tracklets to a player identity.

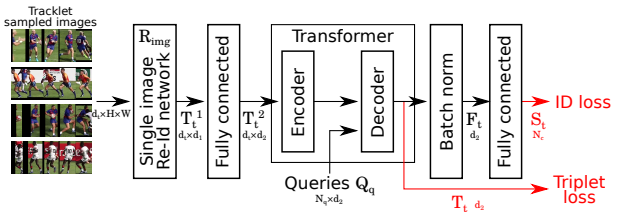


Figure 3. Architecture of the tracklet classification network. $R_{img}$ extracts re-ID features $T_t^1$ from the tracklet images. They are combined by the transformer to generate a single tracklet re-ID vector $F_t$. The model is trained by ID loss and triplet loss.

for every players in a team with their semi-supervised approach. Senocak et al. [43] use 2500 cropped images for each player to train their re-ID network. Teket and Yetik [44] use a training dataset that contains 30 to 1000 images per player. In this paper, by asking the user to annotate tracklets, we aim to demonstrate that it is possible to produce meaningful player re-ID results for a rugby sevens full game with only 6 annotations per player.

## 3. Proposed method

### 3.1. Overview

We propose a new method to track the $N_p$ players of a team in a video with a single moving view of a game. The first step of our method generates $N_t$ tracklets we qualify as non-ambiguous because they contain a single identity. For this purpose, bounding boxes around persons are detected and associated across frames automatically. The user can then provide few identity annotations to some of the generated tracklets thanks to a dedicated interface show on Figure 4. The tracklet re-ID network can then be trained with these annotations. Once the model is trained, classification scores and re-ID features are generated for all the tracklets. This data feeds an algorithm that matches every tracklet to an identity. Once the annotation interface has been updated, the user can then decide to add more annotations to correct the wrong classifications or to stop this incremental learning mechanism if she/he is satisfied by the results. The whole process is depicted on Figure 2.

### 3.2. Tracklet generation

Non-ambiguous tracklets are generated with a tracking by detection paradigm. A Faster R-CNN network [40] with a ResNet-50 backbone [19] trained on the COCO dataset [31] detects all the persons in the video frames. This detector is a well-known model used in several recent work [24, 47]. To generate the tracklets, we use the simple and classic approach described in [4]. Bounding boxes between the previous and the current frames are associated by bipartite matching with an Hungarian algorithm [29]. This matching is performed with a single IoU criteria since the player appearances are later taken into account by our tracklet re-ID model. We also use a Kalman filter [25] to predict the position of an existing track in the current frame.

Each generated tracklet will be later matched to a single identity. We therefore want to avoid as much as possible identity switches inside tracklets. When a tracklet partially occludes an other one, bipartite matching may generate a wrong association. Our algorithm therefore splits the tracklets that intersect since they are considered as ambiguous. If at the current frame, two tracklet bounding boxes have an IoU above a threshold $\mu = 0.5$ these tracklets are terminated and new ones are created. We also filter out tracklets that have a length inferior to $l_{min}$. We indeed consider that they may also be ambiguous by containing several identities in their images. Besides, they do not provide enough diverse data to the tracklet re-ID model.

### 3.3. Incremental learning tracklet classification

The aim of our system is to match tracklets to identities with the fewest possible annotations. This process is done through incremental learning since the user can choose to add more training annotations while the quality of the generated tracklet association is not satisfying. We set the target number of classes $N_c = 1+N_p$. The class zero corresponds to all persons we do not want to track (players from the opponent team, referees, public). Our tracklet re-ID model is mainly composed of a single image re-ID network $R_{img}$ followed by a Transformer [46] as illustrated on Figure 3.

For $R_{img}$, we chose the model described by Luo et al. [35] for its simplicity. It uses a ResNet-50 backbone [19] and has been trained on the generic Market1501 dataset [57]. It takes as input single images at resolution $H \times W$ and outputs player appearance features at dimension $d_1$. We regularly sample $d_t$ images from each tracklet and combine their appearance features to obtain the tracklet features tensor $T_t^1 \in \mathbb{R}^{d_t \times d_1}$. The feature dimension of $T_t^1$ is then reduced to $d_2$ by a fully connected layer to obtain $T_t^2$. This limits the dimension of the features inside the next nodes of

our model in order to train it quickly.

The Transformer in our model then combines the re-ID features of the sampled tracklet images $T_t^2$ to generate a single tracklet re-ID vector $F_t$. Its cross-attention nodes can learn to focus on the most distinctive features across the tracklet sampled frames. It takes as input of the encoder $T_t^2$ and as input of the decoder the $N_q$ queries $Q_q$. Similarly to DETR [6], the queries $Q_q \in \mathbb{R}^{d_2}$ are learned embeddings. Each query learns to specialize on some features of the player identities. However, we do not use any input positional encoding because, since our initial variable length tracklets are resampled to fixed length $d_t$, there are no common temporal link between the features. We found that using 16 encoder layers, one decoder layer and 16 heads in the multi-head attention models was the best set of parameters. At the output of the decoder, a batch norm layer generates the tracklet features $F_t$. For the classification, a fully connected layer computes the classification scores $S_t \in \mathbb{R}^{N_c}$.

Given a tracklet $t$, the $N_{qc}$ queries among $N_q$ that gives the highest classification scores are selected for the backpropagation. The optimized loss is defined by

$$L = L_{ID}(S_t, \hat{S}_t) + \alpha L_{Triplet}(D_{t,p}, D_{t,n}),$$

where $L_{ID}$ is the standard cross entropy loss, $\hat{S}_t$ are the target classification logits, $L_{Triplet}$ is the soft-margin triplet loss [21], $D_{t,p}$ and $D_{t,n}$ are feature distances of positive pairs and negative pairs and $\alpha$ is a binary constant. As described by Luo et al. [35], the idea of combining a classification loss and a triplet loss is to let the model learn more discriminative features $F_t \in \mathbb{R}^{d_2}$. For the triplet loss, we use a batch hard strategy that finds the hardest positive and negative samples.

Once the model has been trained, all tracklets are processed by the model at inference stage to compute the tracklet classification scores $S_t$ and features $F_t$.

## 3.4. Association algorithms

With generated scores $S_t$ and the features $F_t$, we have the needed data to match tracklets to player identities by using an association algorithm. Two alternative methods are investigated.

### 3.4.1 Iterative association

An iteration of the association algorithm consists in selecting the highest score in the matrix of all tracklet scores $S_t$. The highest score represents a matching between the tracklet $t$ and the identity $i$. The algorithm then checks that $t$ can be associated to $i$ by verifying that the tracklets already associated to $i$ do not already appear in the frames where $t$ appears. If the association is possible, $t$ is added to the list of tracklets associated to $i$ and a new iteration of the algorithm is run. When the iterative association is used, we set $\alpha = 0$ during the incremental learning to only optimize the classification scores $S_t$.

### 3.4.2 Matrix factorization association

The second algorithm is inspired from [20]. The authors describe a multi-camera batch people tracking system that assigns tracklets extracted from different views to identities. The input of the algorithm is a tracklet similarity matrix $S$ generated with appearance, motion and localization criteria. A Restricted Non-negative Matrix Factorization (RNMF) algorithm optimizes the identity assignment. The association matrix $A \in \mathbb{R}^{N_t \times N_p}$ is computed thanks to the iterative updating rule given in [10]. We applied the RNMF algorithm to our single view case with $S$ as the sum of an appearance term $\Psi_{app}$ and a localization term $\Psi_{loc}$. The similarity between two tracklets $u$ and $v$ is computed with:

$$S(u,v) = clip(\Psi_{app}(F_u, F_v)) + clip(\Psi_{loc}(B_{ul}, B_{vf}))$$

where $clip(x) = max(min(x;1);0)$.

$\Psi_{app}$ is defined by equation 1.

$$\Psi_{app(F_u,F_v)} = 1 - \frac{1}{\eta_{app}} \cdot d(F_u, F_v) \tag{1}$$

where $d(F_u, F_v)$ is the cosine distance between the feature vectors of the two tracklets and $\eta_{app}$ is the cosine distance threshold above which we consider that $u$ and $v$ belongs to two distinct identities.

$\Psi_{loc}$ is defined by the equation 2. $t_{ul}$ is the end time of the of the first tracklet and $t_{vf}$ is the start time of the second tracklet. $B_{ul}$ and $B_{vf}$ are the corresponding bounding boxes.

$$\Psi_{loc}(B_{ul}, B_{vf}) = \begin{cases} (1+\eta_{loc}) \cdot IoU(B_{ul}, B_{vf}) - \eta_{loc} & \text{if } t_{vf} - t_{ul} \leq \tau \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\eta_{loc}$ and $\tau$ are constant numbers. $\Psi_{loc}$ aims at giving a high similarity scores to two successive tracklets if $B_{ul}$ and $B_{vf}$ have a high IoU. When the RNMF association is used, we set $\alpha = 1$ during the incremental learning.

# 4. Experimental Results

## 4.1. Implementation details

Our system is implemented using the Pytorch framework. The minimum number of frames of a tracklet $l_{min}$ is set to 10. All the tracklets are resampled to $d_t = 10$. Our re-ID network [35] takes as input images of resolution $H = 256$ and $W = 128$. It outputs features at dimension $d_1 = 2048$. Our Transformer network takes input features


Figure 4. Partial screen capture of our semi-interactive annotation interface. Each cell corresponds to one tracklet. Each column corresponds to one player identity, except the zero column that contains all the persons we do not want to track.

at $d_2 = 128$. The number of input queries $N_q$ is set to 32. They are randomly initialized. The number of queries selected for backpropagation $N_{qc}$ is set to 4. It is trained during 120 epochs with an AdamW optimizer, a learning rate of $9 \times 10^{-5}$, a weight decay of $10^{-4}$ and a batch size of 4. The transformer parameters are initialized with Xavier initialization [16]. For the linear layer, He initialization [18] is used. $\eta_{app}$ and $\eta_{loc}$ are experimentally set to 0.35 and 0.43. The time threshold $\tau$ for the localization similarity is set to 0.5 seconds.

Our semi-interactive annotation interface, illustrated on Figure 4, can run on a laptop GPU (Quadro M2000M). It allows the annotator to generate training data for our model by indicating to which player belongs a tracklet. The training time represents about 0.8 second per annotation when $R_{img}$ is frozen and the iterative association is used.

## 4.2. Player tracking on rugby sevens samples

### 4.2.1 Dataset

Rugby sevens is a variant of rugby where two teams of seven players play a game composed of two seven minute halves. It is an Olympic sport since 2016. We annotated a total of 58193 person bounding boxes in the images of three rugby sevens samples of 40 seconds to use them as ground truth for players of both teams, the referees and some people in the public. These samples come from the Argentina / France, France / Chile and France / Kenya games of the 2021 Dubai tournament. They are encoded at a resolution of 1920 by 1080 pixels and a frame rate of 50 frames per seconds. The aim of our experiments is to track players from one of the two teams taking part to the game.

Tracklets were extracted with the method detailed in section 3.2. About 30% of the tracklets have a number of frames superior to $l_{min} = 10$. This represent an average of 346 tracklets per video of 40 seconds. These tracklets have an average length of one second and correspond to about 89% of the detected bounding boxes. We publicly release the tracking ground truth and the generated tracklets at https://kalisteo.cea.fr/index.php/free-resources/.

### 4.2.2 Quantitative results and ablation studies

The annotator selects a number of tracklet examples for each player appearing in the sequence and also for the class 0 (opponent team, referees, public). At each round of annotations, a new user annotation for each player and two user annotations for the class 0 are added on average. As the training of our system is quick, the user can observe the consequences of the added annotations on the classification results and correct the big mistakes for the next round of annotations (for example false positives with high scores).

Once the user annotations have been added, we train the network with the same user annotations and 5 different seeds. We then compute standard MOT metrics [41]: IDF1, MOTA and ID switches. Since our main objective is to correctly identify each player, the IDF1 metric is the most important to observe. MOTA is however key to report the completeness of the tracking bounding boxes for each player. Figure 5 shows the results of our method obtained with four variants. Results are analyzed according to two conditions: $R_{img}$ frozen or trained and with the iterative association algorithm or with the RNMF algorithm. As small tracklets are filtered, our method cannot achieve 100% performance. In order to estimate the upper performance limit, we associate each tracklet to the ground truth. However, since our generated tracklets are not perfect, their association to the ground truth may also be ambiguous, which explains the not null ID switch limits.

**Number of annotations**. For the three video extracts, the more user annotations are provided, the best the MOT metrics are. However, we can observe that above the third round of annotations (about 3.5 annotations per player), the metrics only slightly improve and sometimes slightly deteriorate. This performance threshold can be explained by the difficult tracking conditions of some instants: the players are sometimes highly occluded or very small, there are very few details to identify them and the detection is difficult on complex postures. Some errors are illustrated on Figure 6. From the first to the third round of annotations, with $R_{img}$ frozen and the iterative association algorithm, IDF1 and MOTA metrics increases on average respectively by 11 and 9 p.p. (percentage points) while the ID switches is divided by 5.

**Association algorithm choice**. The global RNMF optimization matches an identity to each tracklet but sometimes generates conflicts and wrong associations. This leads to

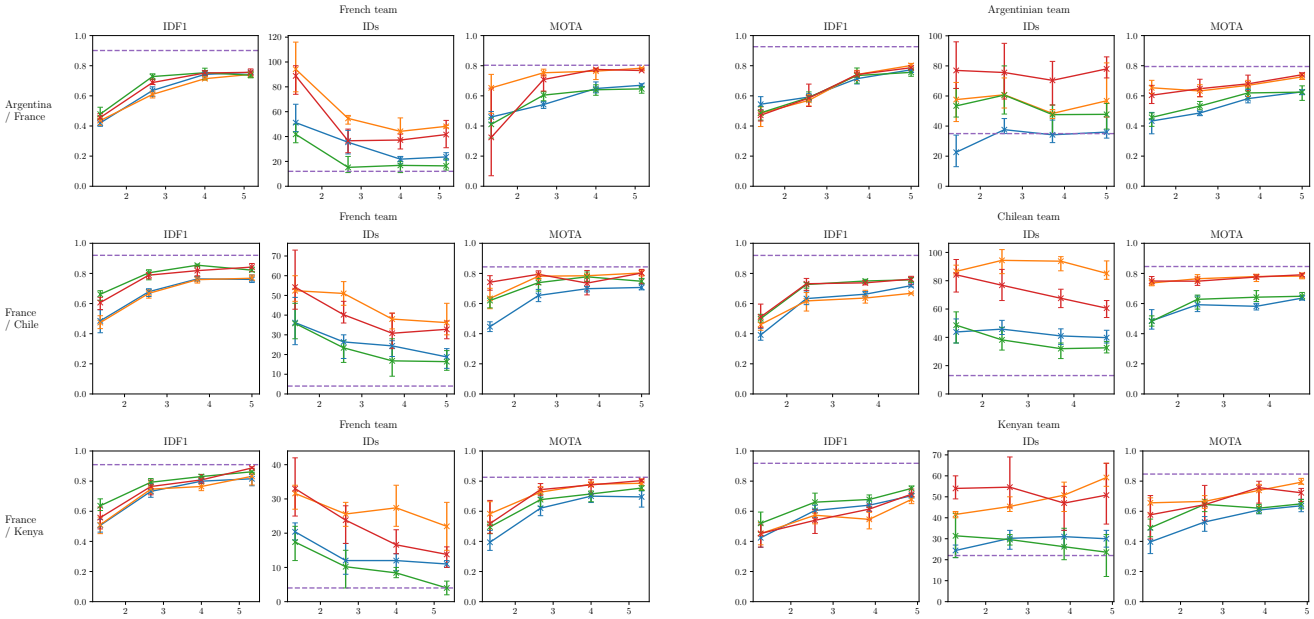


Figure 5. MOT metrics for the tracking of rugby sevens players in 3 videos. The x-axis corresponds to the total number of annotations divided by the number of tracked players. The variation intervals for the 5 seeds and average values are represented. The tested variants are: $R_{img}$ frozen with the iterative association (—), $R_{img}$ frozen with the RNMF association (—), $R_{img}$ trained with the iterative association (—), $R_{img}$ trained with the RNMF association (—) and the ground truth association (- - -).


Figure 6. Illustration of complex situations that lead to missing detections and identification of players (here the players in blue).

better MOTA metrics as more detections are kept than with the simple iterative algorithm. For the third round of annotations, the MOTA metric is increased by 12 p.p. on average when $R_{img}$ is frozen. However, the IDF1 metric is decreased by 1 p.p. and the number of ID switches increases by 25. The iterative association should therefore be preferred to minimize wrong identity associations. The RNMF algorithm however leads to a more complete tracking.

**Training strategy**. Our experiments demonstrate that, even if $R_{img}$ is not fine-tuned with data from the target domain ($R_{img}$ frozen), it is still able, thanks to the Transformer network, to generate relevant features to re-identify the players. For the third round of annotations, the IDF1 and MOTA metrics are respectively on average 75% and 66% with the iterative association algorithm. The best results are however obtained when weights of $R_{img}$ are also updated during training. For the third round of annotations, with the iterative association algorithm, the IDF1 and MOTA metrics are increased respectively by 3 and 2 p.p. The number of ID switches is reduced on average by 3. When the weights of $R_{img}$ are updated, the training time for the 120 epochs significantly increases (from 28 seconds to 25 minutes for 32 annotations) and the system is no longer interactive. Indeed, the number of trainable parameters raises from about 4 to 25 millions. So, the optimal usage is to create the user annotations with $R_{img}$ frozen and once the user is satisfied with the results, restart the training with the same annotations and $R_{img}$ updated to obtain even better results.

#### 4.2.3 Comparison with state of the art multiple person tracking methods

Generic tracking algorithms track all the persons appearing in the video frames. This would include in our case, players of both teams, the referees and the public. Our approach however track players from a single team. This makes the comparison not straightforward. Some approaches have been proposed for the tracking of team sport players with single moving views [24, 34] but the comparison is still not easy since their evaluation datasets are private. We therefore decided to run generic tracking algorithms on our rugby sevens extracts. In order to make a fair comparison with our approach, we manually selected the tracks generated

| Video | Method | IDF1 | IDs | MOTA |
|---|---|---|---|---|
| Argentina / France | ByteTrack [55] | 48.8 | 26 | 49.4 |
| | TWBW [3] | 24.4 | 64 | 40.8 |
| | MOT neur. solv. [5] | 34.0 | 54 | 33.9 |
| | Ours | **76.8** | **17** | **64.6** |
| France / Chile | ByteTrack [55] | 54.9 | 23 | 64.4 |
| | TWBW [3] | 22.7 | 74 | 28.4 |
| | MOT neur. solv. [5] | 29.6 | 53 | 40.2 |
| | Ours | **84.3** | **21** | **75.4** |
| France / Kenya | ByteTrack [55] | 60.3 | 14 | 64.0 |
| | TWBW [3] | 30.6 | 44 | 45.0 |
| | MOT neur. solv. [5] | 48.0 | 26 | 61.1 |
| | Ours | **82.2** | **7** | **70.1** |

Table 1. MOT metrics for the tracking of the rugby sevens French team players.

by these algorithms that are associated, even partially, with players from the French team. We tested two online methods, TWBW tracker [3] and ByteTrack [55], with their detections. ByteTrack achieves a very high performance on the MOT 2017 challenge [38]. We also tested an offline method, the MOT neural solver [5], with our detections. The results are presented in Table 1. With the limitations mentioned above, the metrics shows significantly lower performances for generic trackers. This is probably due to difficulties to handle correctly the occlusions and the players entering or leaving the view field. It therefore justifies our usage of a closed identity gallery with few annotations to learn the player appearances. Compared to ByteTrack [55], the IDF1 metric is increased on average by 26 p.p.

### 4.3. Evaluation of player identification on a full rugby sevens game

Our system aims to track and identify players on a full game. Yet, a human-annotated tracking ground truth for a full game would be costly to generate. We therefore decided to evaluate the detection and re-ID performance of our approach on 32 frames regularly sampled in the France / Kenya game and focus on the French players. With players changes, 12 French players in total participated to this game. The ground truth represents 128 players bounding boxes. For each experiment, we trained the model with 5 different seeds using the same 70 annotations (about 6 per player). Results are shown in Table 2. The best total detection and identification performance (53.6%) is obtained when the $R_{img}$ is trained and the RNMF association algorithm is used. A significant number of French players are not detected or correctly identified. This happens when the players on the back are only visible on few pixels or when some players occlude others. Nevertheless, the total recall goes up to 67.9% for the bounding boxes with an area superior to the average area of all the ground truth bounding boxes (25214 pixels). This demonstrates that when the players are sufficiently visible, our system is able to track them during a full match with few annotations.

| $R_{img}$ | Assoc. | Det. recall | Team class. recall | Id. class. recall | Total recall |
|---|---|---|---|---|---|
| All detected bounding boxes | | | | | |
| frozen | iter. | 75.8 | 58.4±2.1 | 73.8±4.5 | 32.7±2.4 |
| frozen | RNMF | | 74.6±2.5 | 60.9±6.5 | 34.5±4.6 |
| trained | iter. | | 75.9±3.9 | **84.0±3.4** | 48.3±3.0 |
| trained | RNMF | | **89.1±2.0** | 79.4±2.6 | **53.6±1.8** |
| Big detected bounding boxes (area superior to 25214 pixels) | | | | | |
| frozen | iter. | 89.7 | 60.8±2.2 | 77.3±6.8 | 42.1±3.1 |
| frozen | RNMF | | 72.3±2.2 | 66.4±5.0 | 43.1±4.4 |
| trained | iter. | | 76.2±3.5 | **87.4±5.2** | 59.7±4.2 |
| trained | RNMF | | **90.8±0.9** | 83.5±3.4 | **67.9±2.6** |

Table 2. French player detection and classification results on 32 frames of the France / Kenya game for 5 different seeds. Average values and standard deviations are provided. The detection recall corresponds to the number of players detected. The team classification recall corresponds to the number of players classified as French among the detected players. The identity classification recall corresponds to the number of correctly identified players among the players classified as French. The total recall is the product of all the previous columns and represents the complete performance of our system.

## 5. Conclusion

In this paper, we proposed a new method to track team sport players with few user annotations. We demonstrated the performance of our approach on a rugby sevens dataset that we publicly release. We also showed that our method can track rugby sevens players during a full match with the annotation of only 6 few seconds length tracklets per player if they are observable with a minimal resolution. To our knowledge, no previous work on tracking of rugby players has been published. As future work, we would like to improve the detection of small and partially occluded players. Since our approach can be applied to any team sport, we would like to test it on other sports such as basketball. We also believe that the user annotation step would be sped up if an active learning process could smartly suggests tracklets to annotate.

## 6. Acknowledgments

This work benefited from a government grant managed by the French National Research Agency under the future investment program (ANR-19-STHP-0006) and the FactoryIA supercomputer financially supported by the Ile-de-France Regional Council. The videos of our Rugby Sevens dataset are the courtesy of World Rugby. We would also like to thank Jérôme Daret, Jean-Baptiste Pascal and Julien Piscione from the French Rugby Federation for making this work possible.

## References


[1] Sermetcan Baysal and Pınar Duygulu. Sentioscope: a soccer player tracking system using model field particles. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(7):1350–1362, 2015. 3

[2] Jerome Berclaz, Francois Fleuret, Engin Turetken, and Pascal Fua. Multiple object tracking using k-shortest paths optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1806–1819, 2011. 2, 3

[3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *IEEE International Conference on Computer Vision*, pages 941–951. IEEE, 2019. 8

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing*, pages 3464–3468. IEEE, 2016. 2, 4

[5] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6247–6257. IEEE, 2020. 2, 3, 8

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 5

[7] Alvin Chan, Martin D Levine, and Mehrsan Javan. Player identification in hockey broadcast videos. *Expert Systems with Applications*, 165:113891, 2021. 3

[8] Damien Delannay, Nicolas Danhier, and Christophe De Vleeschouwer. Detection and recognition of sports (wo) men from multiple views. In *2009 Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–7. IEEE, 2009. 2

[9] Adrien Deliege, Anthony Cioppa, Silvio Giancola, Meisam J Seikavandi, Jacob V Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B Moeslund, and Marc Van Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4508–4519. IEEE, 2021. 2

[10] Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2008. 5

[11] Tiziana D'Orazio, Marco Leo, Nicola Mosca, Paolo Spagnolo, and Pier Luigi Mazzeo. A semi-automatic system for ground truth generation of soccer video sequences. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 559–564. IEEE, 2009. 2

[12] Tiziana D'Orazio, Marco Leo, Paolo Spagnolo, Pier Luigi Mazzeo, Nicola Mosca, Massimiliano Nitti, and Arcangelo Distante. An investigation into the feasibility of real-time soccer offside detection from a multiple camera system. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(12):1804–1818, 2009. 3

[13] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 3

[14] Sebastian Gerke, Antje Linnemann, and Karsten Müller. Soccer player recognition using spatial constellation features and jersey number recognition. *Computer Vision and Image Understanding*, 159:105–115, 2017. 3

[15] Sebastian Gerke, Karsten Muller, and Ralf Schafer. Soccer jersey number recognition using convolutional neural networks. In *IEEE International Conference on Computer Vision Workshops*, pages 17–24. IEEE, 2015. 3

[16] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *The Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 6

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015. 6

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 4

[20] Yuhang He, Xing Wei, Xiaopeng Hong, Weiwei Shi, and Yihong Gong. Multi-target multi-camera tracking by tracklet-to-target assignment. *IEEE Transactions on Image Processing*, 29:5191–5205, 2020. 5

[21] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 3

[23] Andrea Hornakova, Roberto Henschel, Bodo Rosenhahn, and Paul Swoboda. Lifted disjoint paths with application in multiple object tracking. In *International Conference on Machine Learning*, pages 4364–4375. PMLR, 2020. 2

[24] Samuel Hurault, Coloma Ballester, and Gloria Haro. Self-supervised small soccer player detection and tracking. In *The 3rd International Workshop on Multimedia Content Analysis in Sports*, pages 9–18, 2020. 2, 3, 4, 7

[25] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960. 2, 4

[26] Seyed Hossein Khatoonabadi and Mohammad Rahmati. Automatic soccer players tracking in goal scenes by camera motion elimination. *Image and Vision Computing*, 27(4):469–479, 2009. 3

[27] Alireza Khotanzad and Yaw Hua Hong. Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, 1990. 3

[28] Longteng Kong, Mengxiao Zhu, Nan Ran, Qingjie Liu, and Rui He. Online multiple athlete tracking with pose-based long-term temporal dependencies. *Sensors*, 21(1):197, 2021. 3

[29] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 4

[30] Gen Li, Shikun Xu, Xiang Liu, Lei Li, and Changhu Wang. Jersey number recognition with semi-supervised spatial transformer network. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1783–1790, 2018. 3

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 4

[32] Hengyue Liu and Bir Bhanu. Pose-guided r-cnn for jersey number recognition in sports. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[33] Jia Liu, Xiaofeng Tong, Wenlong Li, Tao Wang, Yimin Zhang, and Hongqi Wang. Automatic player detection, labeling and tracking in broadcast soccer video. *Pattern Recognition Letters*, 30(2):103–113, 2009. 3

[34] Wei-Lwun Lu, Jo-Anne Ting, James J Little, and Kevin P Murphy. Learning to track and identify players from broadcast sports videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1704–1716, 2013. 2, 3, 7

[35] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4, 5

[36] Mehrtash Manafifard, Hamid Ebadi, and H Abrishami Moghaddam. A survey on player tracking in soccer videos. *Computer Vision and Image Understanding*, 159:19–46, 2017. 3

[37] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021. 3

[38] 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. 2, 8

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 3

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015. 2, 3, 4

[41] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*, pages 17–35. Springer, 2016. 6

[42] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 3

[43] Arda Senocak, Tae-Hyun Oh, Junsik Kim, and In So Kweon. Part-based player identification using deep convolutional representation and multi-scale pooling. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1732–1739, 2018. 3, 4

[44] Osman Murat Teket and Imam Samil Yetik. A fast deep learning based approach for basketball video analysis. In *4th International Conference on Vision, Image and Signal Processing*, pages 1–6, 2020. 3, 4

[45] Rajkumar Theagarajan and Bir Bhanu. An automated system for generating tactical performance statistics for individual soccer players from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2):632–646, 2020. 3

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 4

[47] Kanav Vats, Pascale Walters, Mehrnaz Fani, David A Clausi, and John Zelek. Player tracking and identification in ice hockey. *arXiv preprint arXiv:2110.03090*, 2021. 2, 3, 4

[48] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7942–7951, 2019. 3

[49] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE International Conference on Image Processing*, pages 3645–3649. IEEE, 2017. 2, 3

[50] Junliang Xing, Haizhou Ai, Liwei Liu, and Shihong Lao. Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling. *IEEE Transactions on Image Processing*, 20(6):1652–1667, 2010. 3

[51] Yukun Yang, Ruiheng Zhang, Wanneng Wu, Yu Peng, and Min Xu. Multi-camera sports players 3d localization with identification reasoning. In *25th International Conference on Pattern Recognition*, pages 4497–4504. IEEE, 2021. 3

[52] Qixiang Ye, Qingming Huang, Shuqiang Jiang, Yang Liu, and Wen Gao. Jersey number detection in sports video for athlete identification. In *Visual Communications and Image Processing*, volume 5960, page 59604P. International Society for Optics and Photonics, 2005. 3

[53] Li Zhang, Yuan Li, and Ramakant Nevatia. Global data association for multi-object tracking using network flows. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[54] Ruiheng Zhang, Lingxiang Wu, Yukun Yang, Wanneng Wu, Yueqiang Chen, and Min Xu. Multi-camera multi-player tracking with deep player identification in sports video. *Pattern Recognition*, 102:107260, 2020. 2, 3

[55] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 2, 8

[56] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888*, 2020. 3

[57] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 4