Re-Identification for Multi-Person Tracking

Vladimir Somers

Keemotion - UCL - EPFL

06/07/2020

Summary

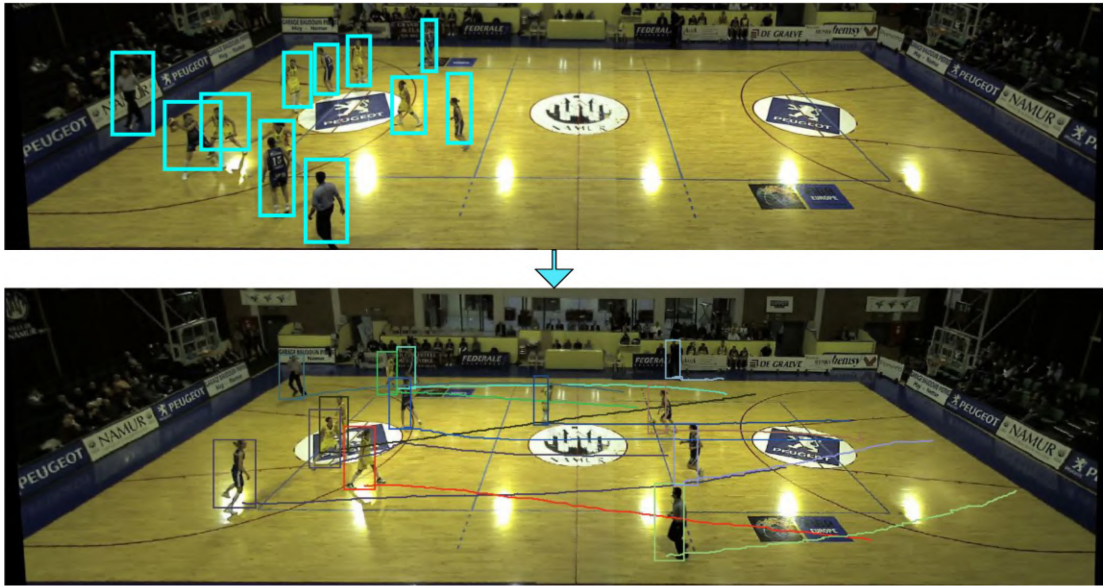
1. **Long term Multi-Object Tracking** and **obstacles**
2. **Visual affinity** estimation for tracklets/detections
3. **Person Re-Identification** models for visual affinity estimation
4. Differences between **traditional Re-ID** and **Re-ID for MOT**
5. **Questions** session

Context: Sport player identification and tracking

Identifying players **during the whole game** by telling where each player is located at each frame.

Identification is mandatory for **personalized content**:

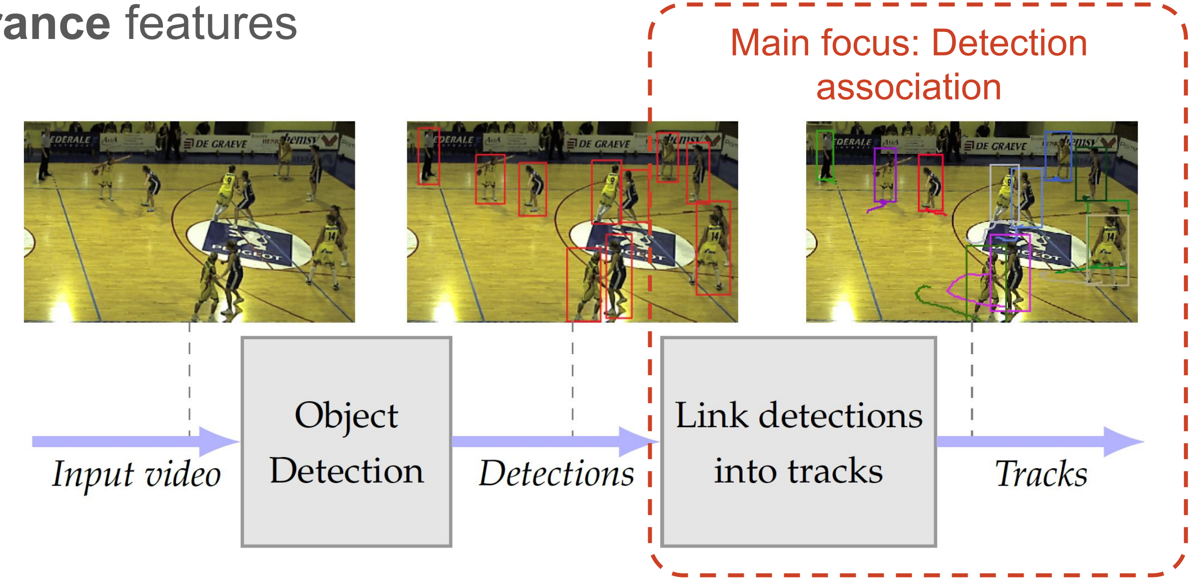
- Individual **statistics**
- Personalized game **highlights**



Multi-Object Tracking (MOT)

Tracking-by-detection is the mainstream approach:

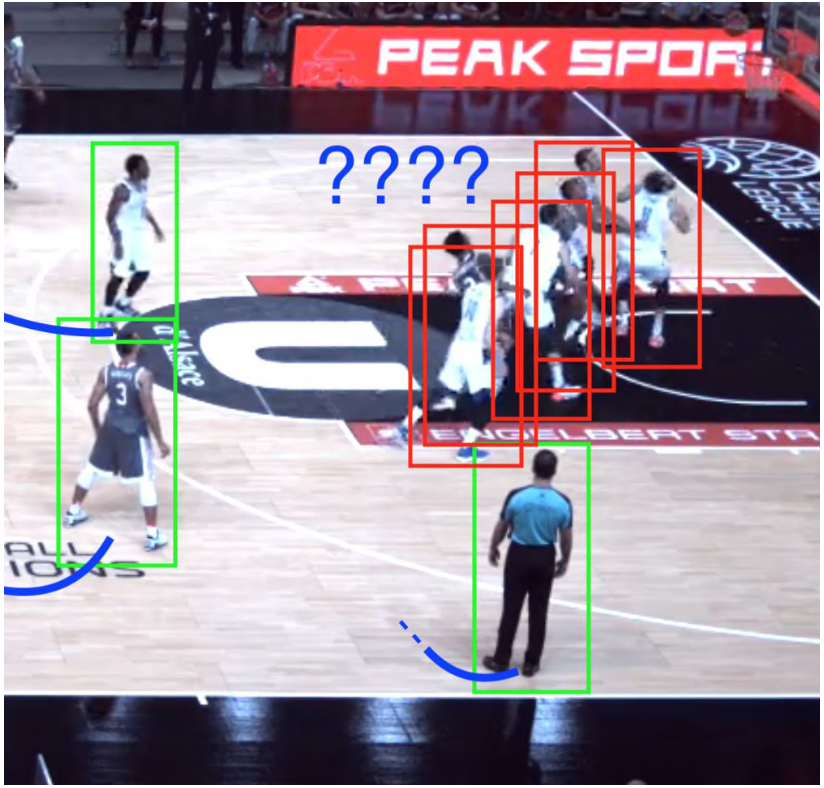
1. **Detection generation** with an object detector
2. **Detection association** over **adjacent frames** using **spatio-temporal** and **appearance** features



Adjacent association fails at long term MOT

Detection association over **adjacent frames** can be :

- Easy :
  - Isolated players
  - Discriminative appearance feature
- Ambiguous or impossible :
  - Occlusions
  - Similar appearance
  - Sporadicity* of discriminative appearance features

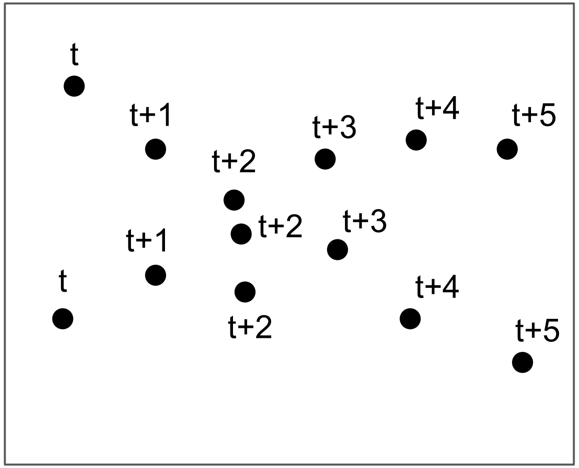


*appearing or happening at irregular intervals in time

From short to long term tracking

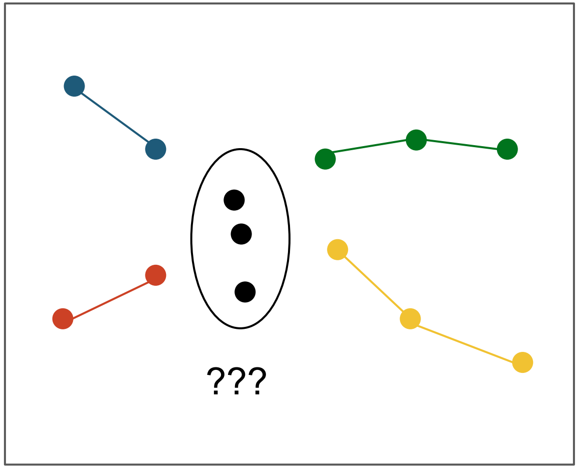
Need to compute a **tracklet affinity metric** to drive the association process.

Detections



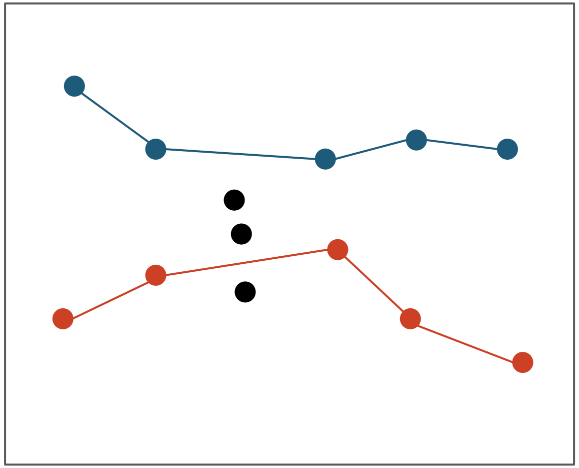
- Detections generated for **successive frames**

Short Tracklets



- **Non ambiguous** adjacent frames association
- **Existing solutions** to produce short tracklet

Long Tracks

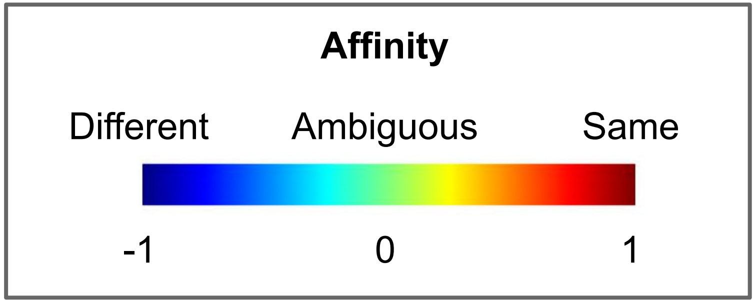


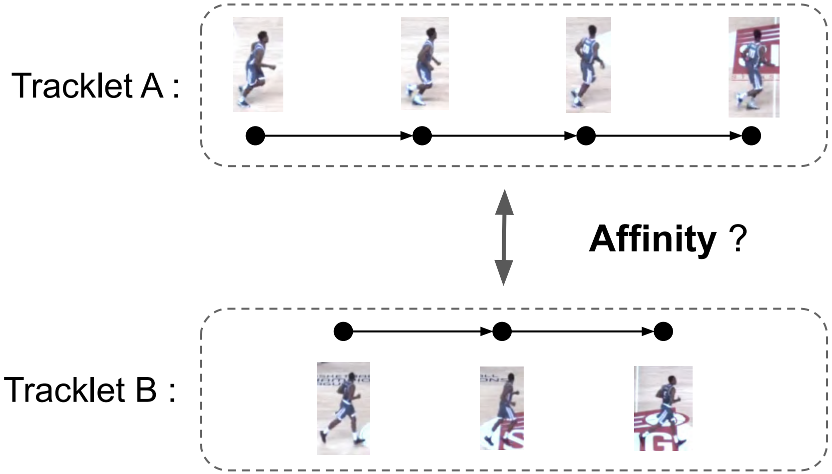
- Tracklets **Re-Identification** with **bridges**
- Avoid **Identity Switches**
- Core problem

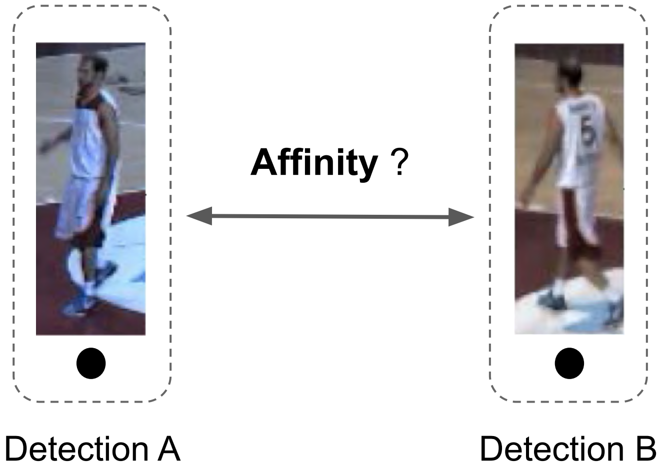
Tracklets/detections affinity

Affinity is the **identity similarity score.**

It's based on **spatio-temporal** and/or **appearance** features.







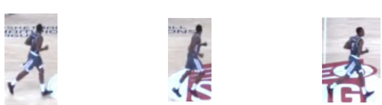
Tracklets/detections affinity

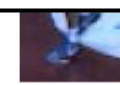
Affinity

Affinity is the **identity similarity score**

THESIS CORE GOAL

Tracklet B :





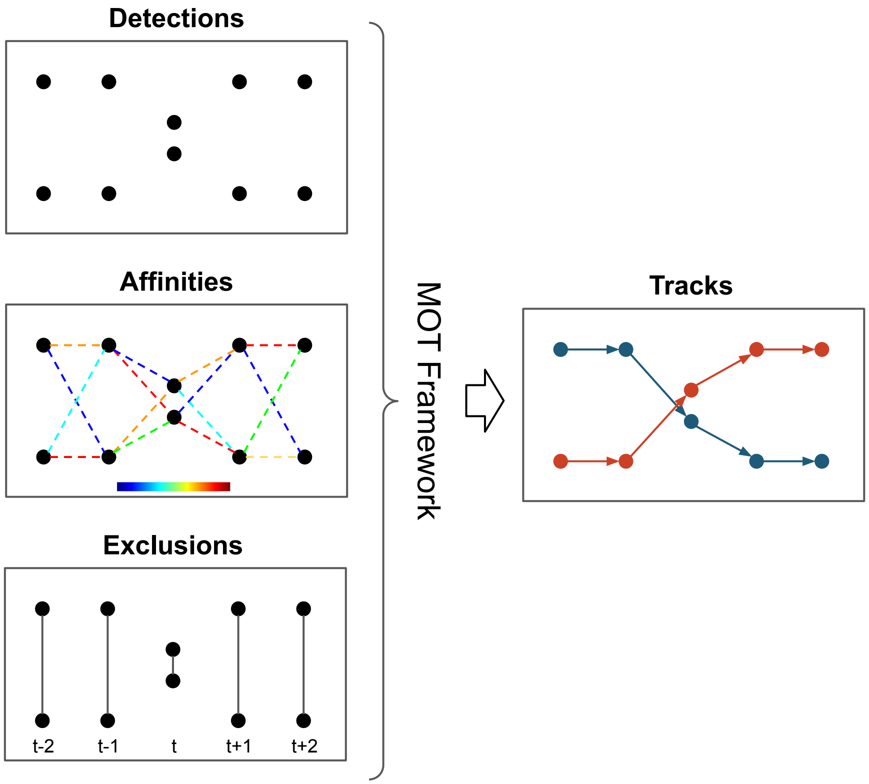
Detection A

Detection B

Graph based formalism for tracks optimization

There are existing **graph-based MOT frameworks** to jointly optimize multiple players **tracks** based on detections/tracklets pairwise **affinities** and **exclusions** graphs

- Amit Kumar, K. C., Delannay, D., & Vleeschouwer, C. De. (2016). ***Iterative hypothesis testing*** *for multi-object tracking in presence of features with variable reliability*.
- Amit Kumar, K. C., Jacques, L., & De Vleeschouwer, C. (2017). *Discriminative and Efficient* ***Label Propagation*** *on Complementary Graphs for Multi-Object Tracking*.

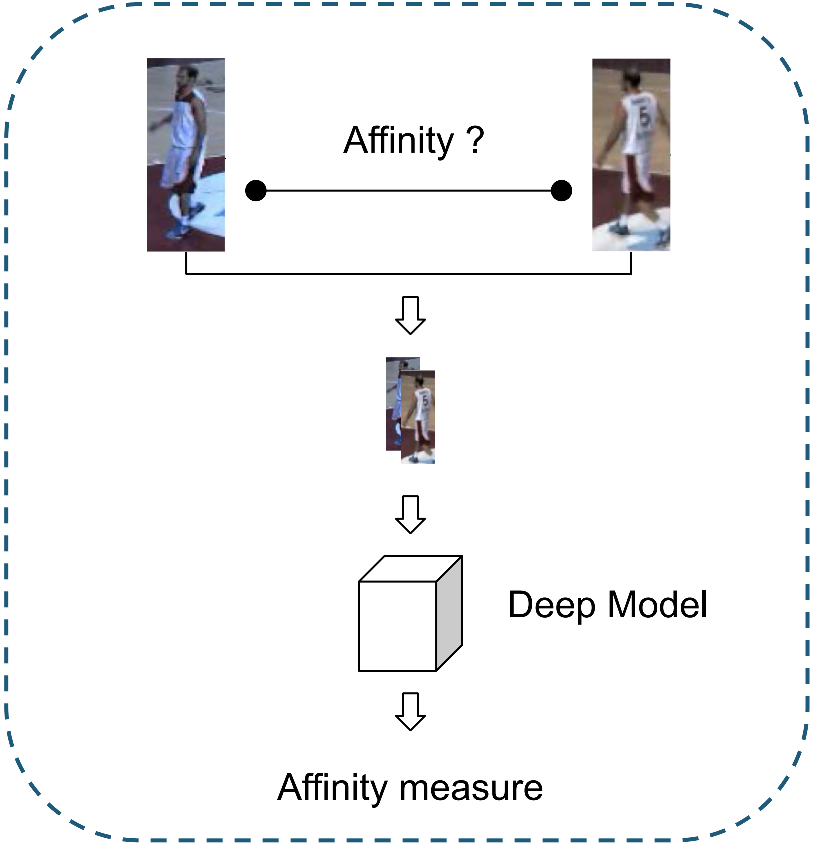


How to estimate detections visual affinity?

Use modern **Deep CNN models**

First approach :

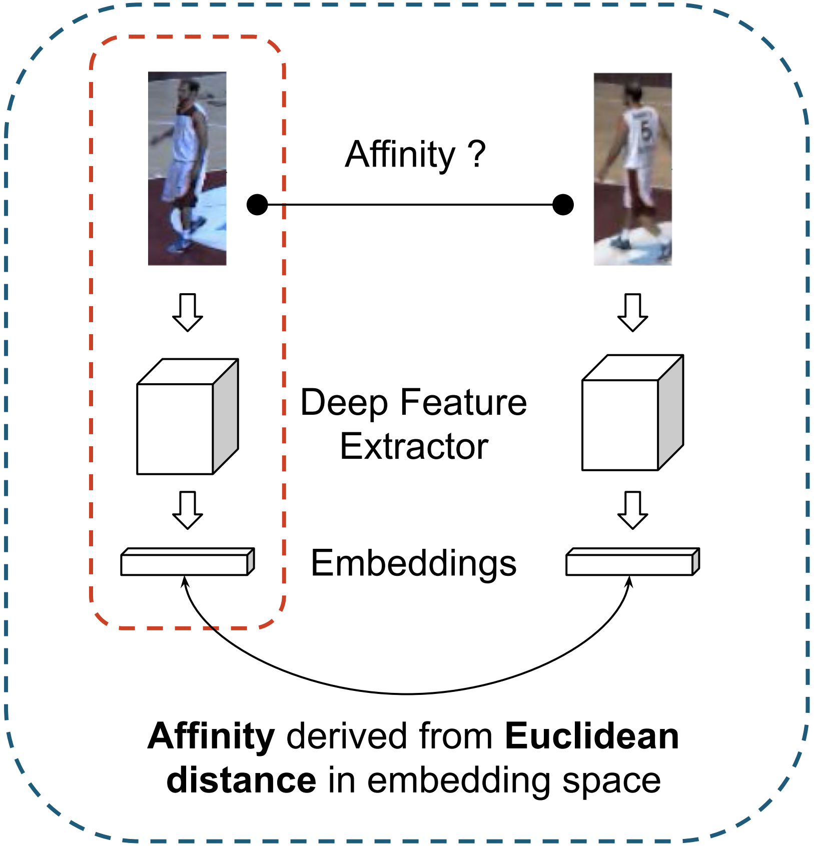
- Direct pairwise affinity inference using a deep model
- Not convenient: **n** detections -> **$O(n^2)$** model inferences



How to estimate detections affinity?

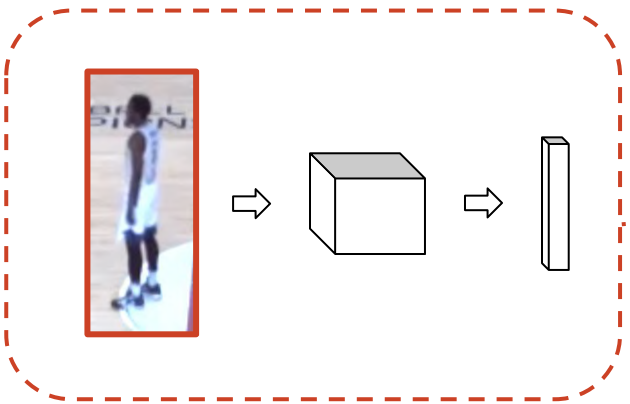
Second approach :

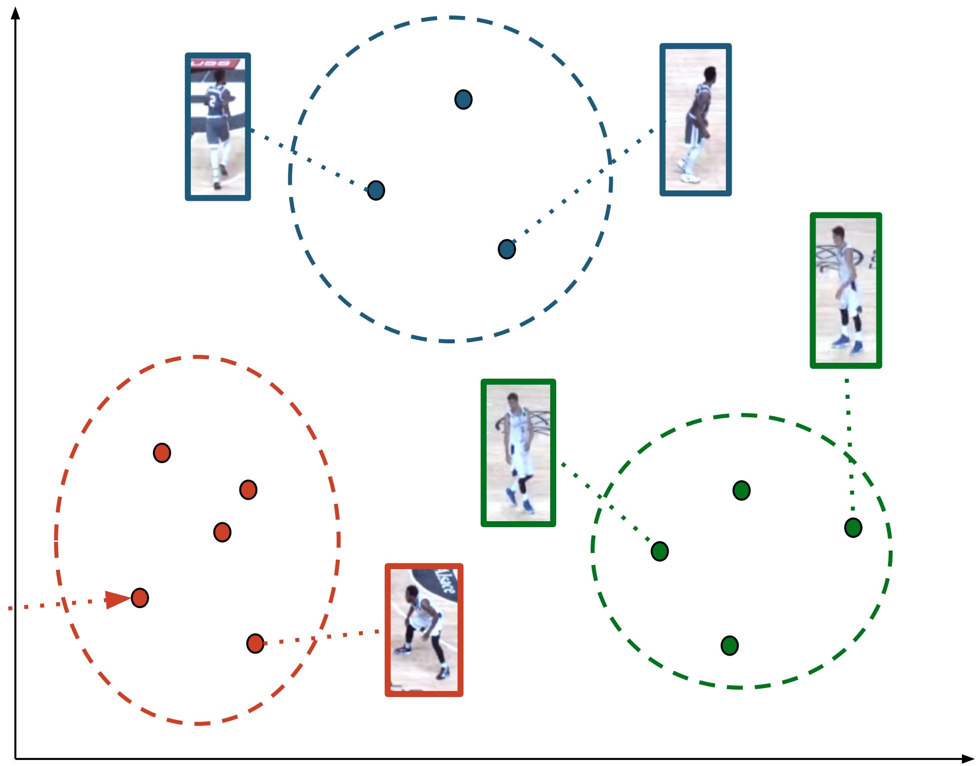
- Deep **feature extractor**
- Detections projected to **embedding space**
- **Affinity** derived from **Euclidean distance**
- **n** model inferences for **n** detections
- **Representation learning**



Representation Learning

- Detections with **same identity** must be **close to each other** in the **embedding space**





2D Embedding Space

How to train a model for Representation Learning?

- Good embeddings = **discriminative features**
- Two popular **training supervision** approaches

Representation Learning:
Training **Supervision**

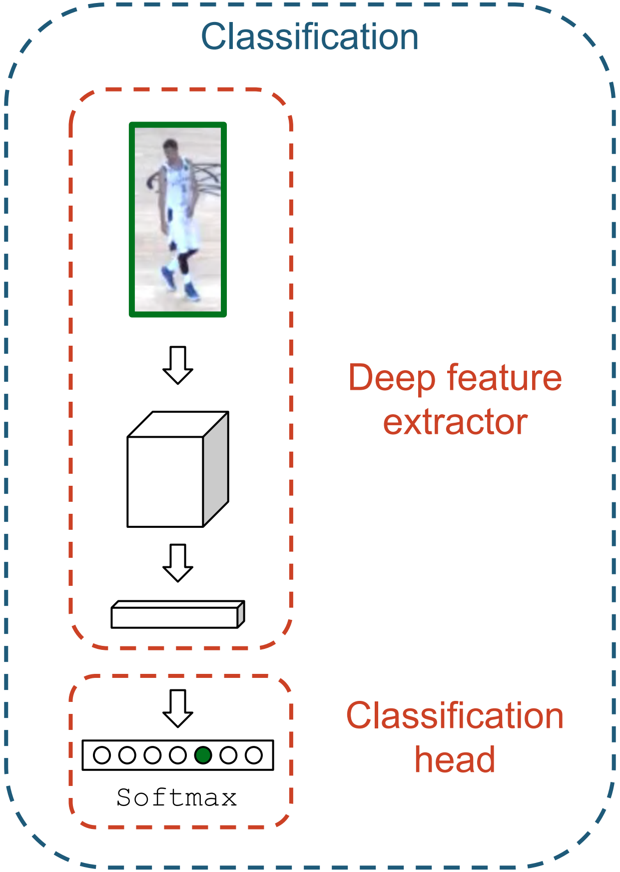
Classification
Cross entropy loss

Distance Metric Learning
Triplet loss

How to train a model for Representation Learning?

Classification Supervision:

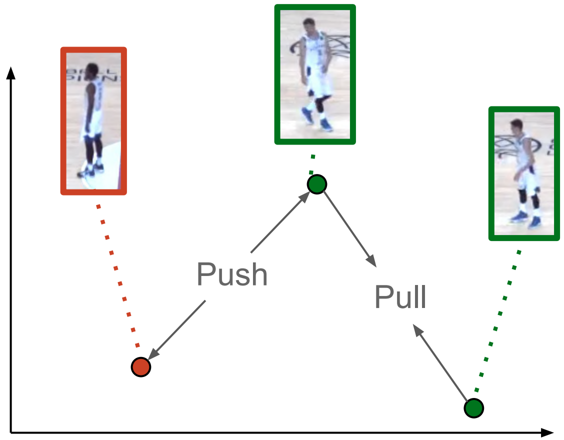
- One **label** for each **identity** in training set
- **Softmax activation** on last layer
- **Cross entropy loss**
- Drop **classification head** at test time



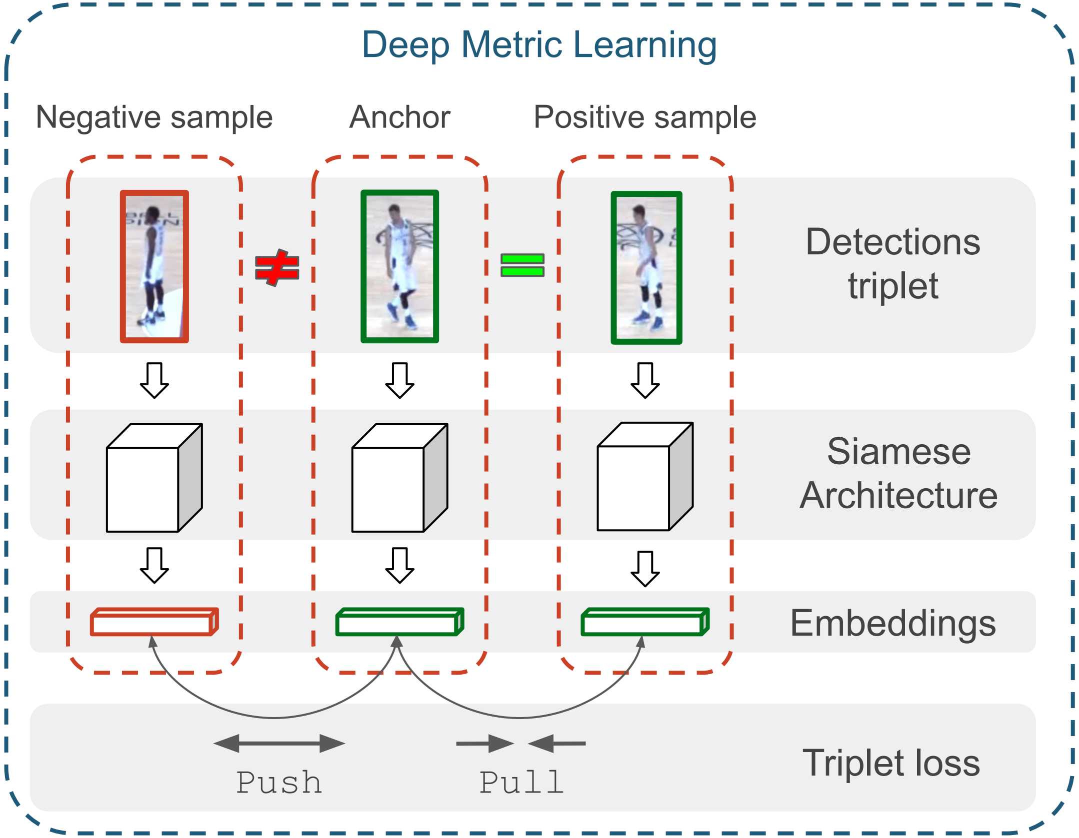
How to train a model for Representation Learning?

Metric Learning Supervision:

- **Triplet loss**
- Euclidean distance in embedding space



Embedding Space

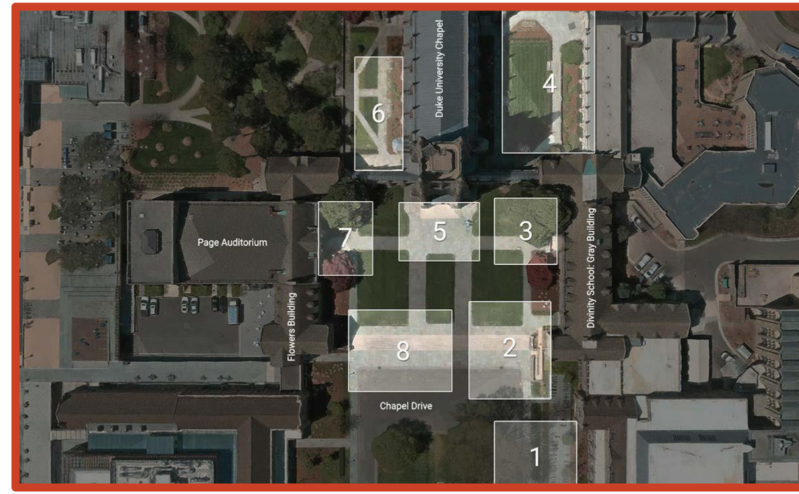


Person Re-Identification as described in literature

Given a person of interest (**query**) find other occurrences of that person among a set of candidates (**gallery**)

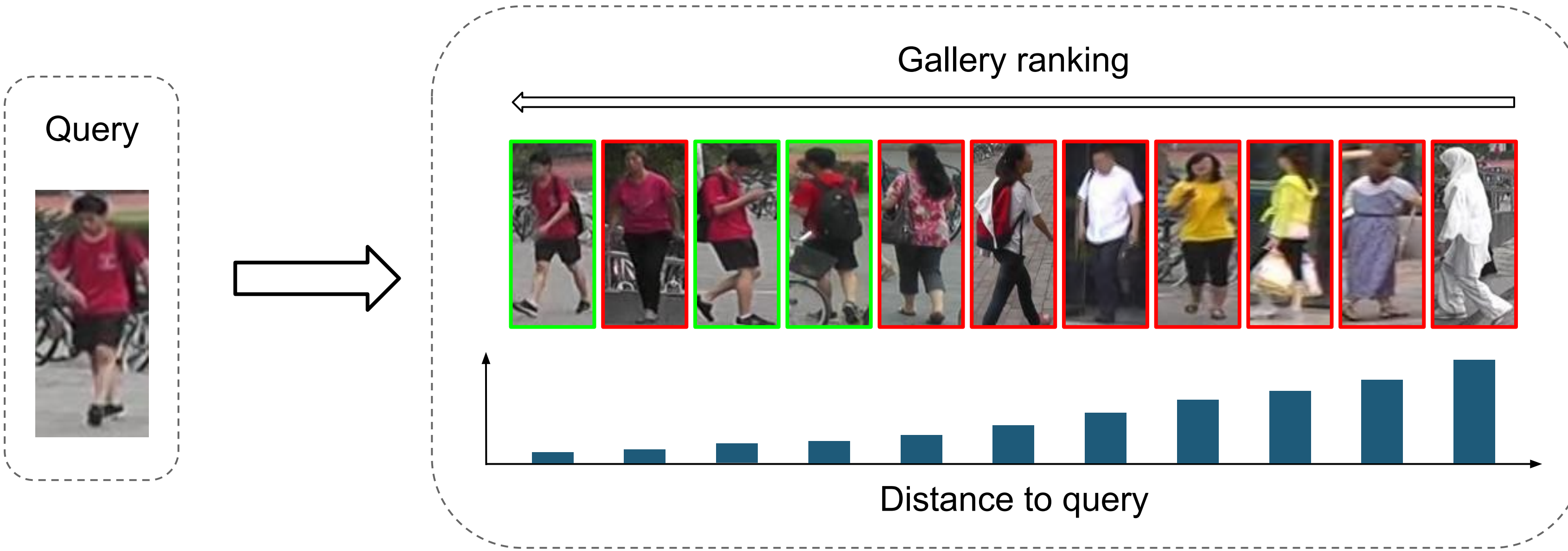
Person Re-Identification

- Multiple surveillance cameras
- Bounding box detections
- Human vignettes gallery



Person Re-Identification objective

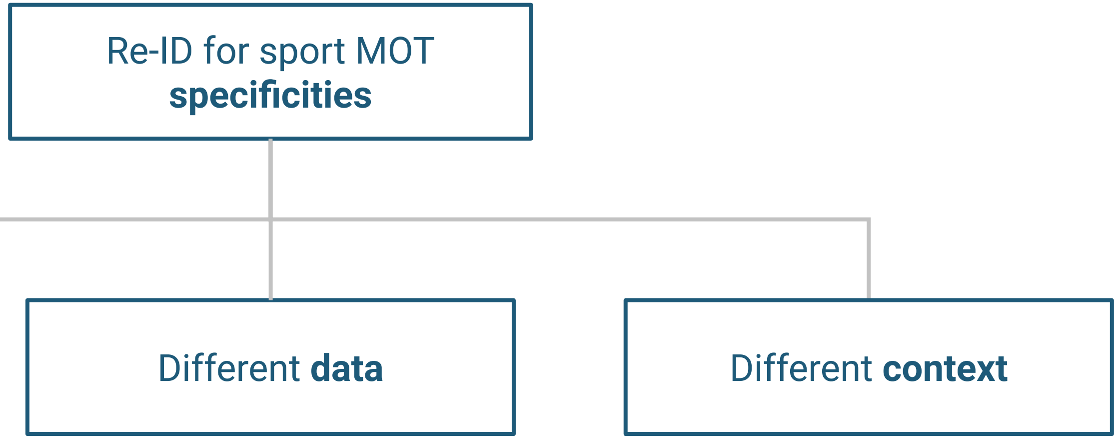
Gallery images are ranked according to their distance to the **query** image



Re-ID for sport MOT vs traditional Re-ID

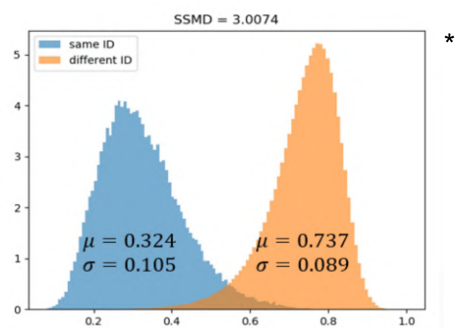
Subtle but important **differences** between **Re-ID for sport MOT** and **traditional Re-ID**

➡ Exploit the **specificities** of this context to build performant Re-ID models



Re-ID for MOT vs traditional Re-ID

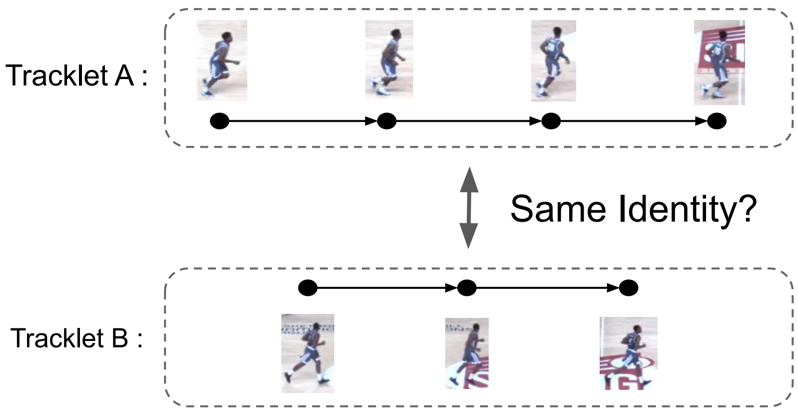
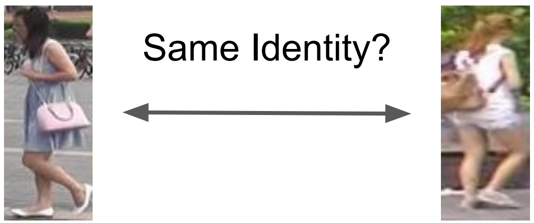
Different **challenges** :

	Re-ID for MOT	**Traditional Re-ID**
Objectives	Tell if pair of images from same/different identity	Find N best match in a gallery
Performance metrics		

* Bastien, N. (2020). *A comparative analysis of deep Re-Identification models for matching pairs of identities.*

Re-ID for MOT vs traditional Re-ID

Different **input data** for re-identification :

Re-ID for MOT	**Traditional Re-ID**
Tracklets comparison	**Image - video** comparison
	

Re-ID for **sport** MOT vs traditional Re-ID

Different **contexts** :

Re-ID for sport MOT	**Traditional Re-ID**
Single camera view Indoor sport pitch No luminosity variation **Similar appearance** and clothes **Sporadicity** of **discriminative appearance features**	Multiple non overlapping surveillance camera Outdoor street view Luminosity, angle and image quality variation Big appearance dissimilarity Discriminative appearance features always available
	

Recap and Questions

Tracklet **visual affinity estimation** using deep **person re-identification models** for solving **long term multi-person tracking** in a **team sport** context.