

Deep Learning for Visual Tracking

T. Chateau

Credits:
Marvasti-Zadeh, Seyed Mojtaba, Li Cheng, Hossein Ghanei-Yakhdan, et Shohreh Kasaei. « Deep Learning for Visual Tracking: A Comprehensive Survey ». CoRR abs/1912.00535 (2019). <http://arxiv.org/abs/1912.00535>.

T. Chateau

1

Deep Learning for Visual Tracking

What is Visual Tracking? From single view single object



2

Deep Learning for Visual Tracking

What is Visual Tracking? To multi view Multi-non rigid-objects



3

Deep Learning for Visual Tracking

What is Visual Tracking?

State Vector

The dynamic configuration of the the tracked object at time k is modelled by a State vector denoted:

$$\mathbf{x}_k$$

State Sequence

The state sequence is given by the set (sequence) of State vectors, denoted:

$$\mathbf{X} \doteq \{\mathbf{x}_k\}_{k=1,\dots,K}$$

Observation

Observation: $\mathbf{Z} \doteq \{\mathbf{z}_k\}_{k=1,\dots,K}$

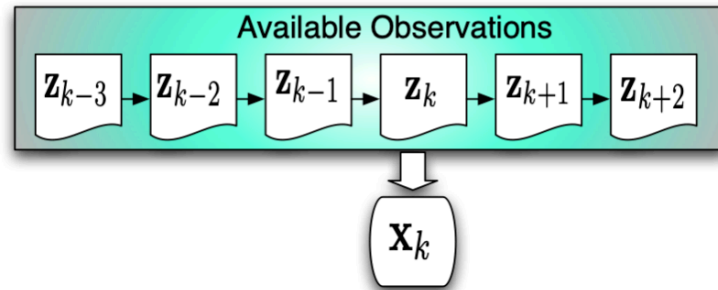
4

Deep Learning for Visual Tracking

Off-line Tracking (Deferred Tracking)

Estimation of the state x_k uses the entire observation sequence

$$\mathbf{Z} \doteq \{\mathbf{z}_k\}_{k=1,\dots,K}$$



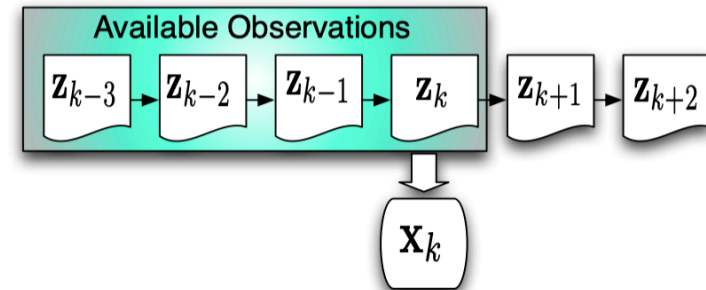
5

Deep Learning for Visual Tracking

On-line Tracking

Estimation of the state x_k uses the current and past observation:

$$\mathbf{z}_{0:k}$$



6

Deep Learning for Visual Tracking

Why is Visual Tracking difficult?

Hidden State

The state \mathbf{X} is a **hidden state** and must be deduced from observation

Tracking Challenges

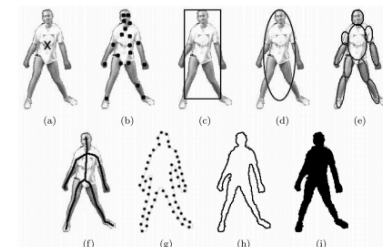
- **Object Modeling**: how to define what an object is in terms that can be interpreted by a computer ?
- **Appearance Change**: The observation of an object changes according to many parameters (illumination conditions, occlusions, shape variation...)
- **Kinematic Modelling**: How to inject priors on object kinematic and interactions between objects.

7

Deep Learning for Visual Tracking

Why is Visual Tracking difficult? (Object representation)

- Object approximation:
 - Segmentation / Polygonal approximation
 - Bounding ellipse/box
 - Position only



- Goal: Measure affinity

Image from A. Yilmaz et. al : Object tracking: A survey. ACM Computing Surveys, 2008

8

Deep Learning for Visual Tracking

Why is Visual Tracking difficult? (Appearance change)

Variation des points de vue



Conditions de luminosité



Deep Learning for Visual Tracking

Why is Visual Tracking difficult? (Appearance change)

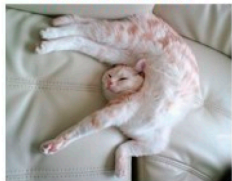
scale variation



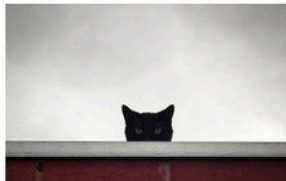
Deep Learning for Visual Tracking

Why is Visual Tracking difficult? (Appearance change)

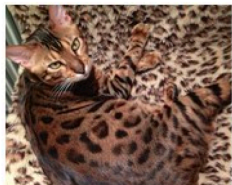
deformable object



occluded object

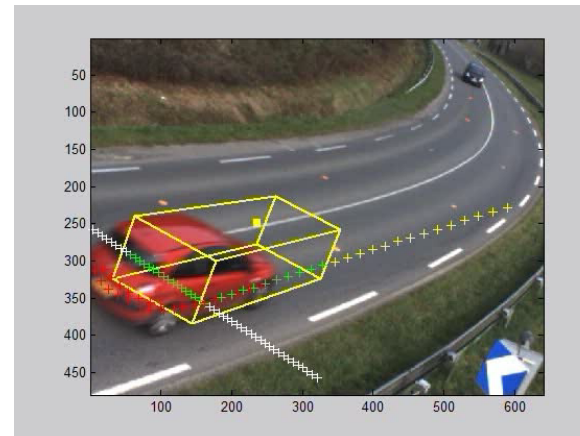


confusion avec l'arrière-plan



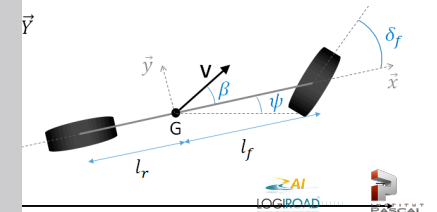
Deep Learning for Visual Tracking

Why is Visual Tracking difficult? (Kinematic modelling)



State Vector:

- 2D position on the ground plane
- 2D motion vector on the ground plane
- Steering angle
- Acceleration



Deep Learning for Visual Tracking

The classical (optimisation) view of tracking (Meanshift)



17

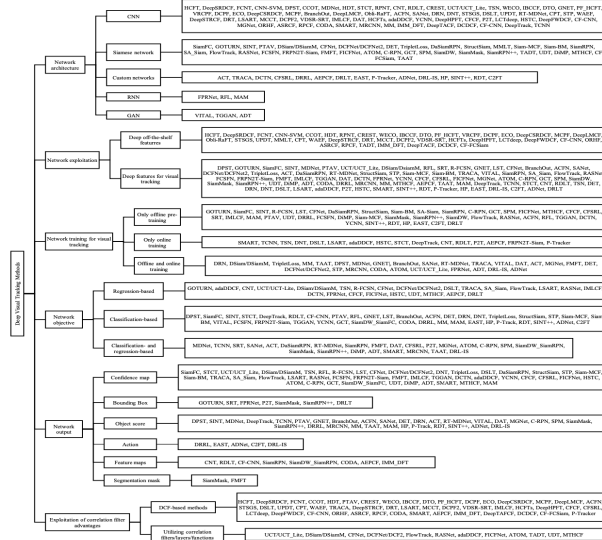
Deep Learning for Visual Tracking

Overview of Visual tracking



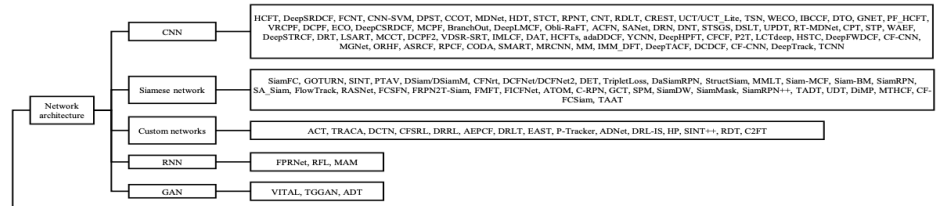
18

Deep Learning for Visual Tracking



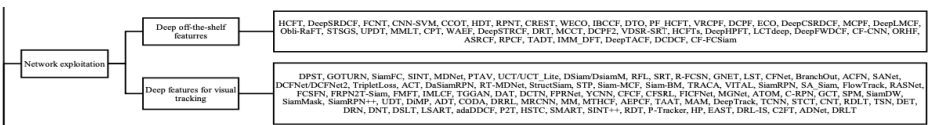
19

Deep Learning for Visual Tracking



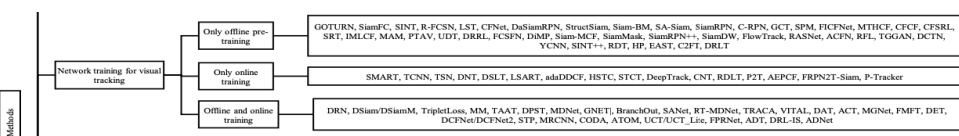
20

Deep Learning for Visual Tracking



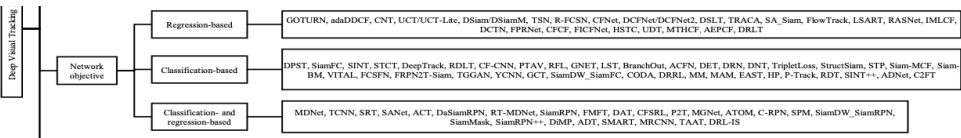
21

Deep Learning for Visual Tracking



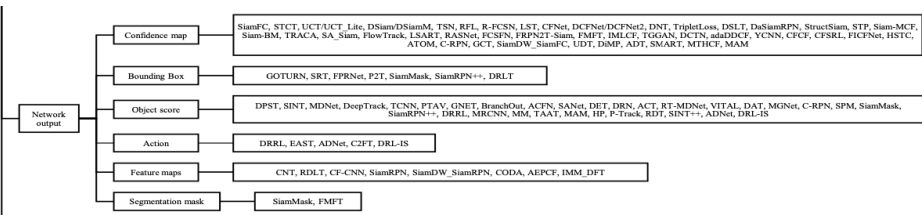
22

Deep Learning for Visual Tracking



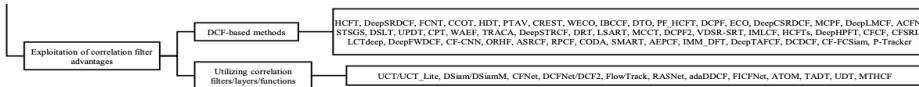
23

Deep Learning for Visual Tracking



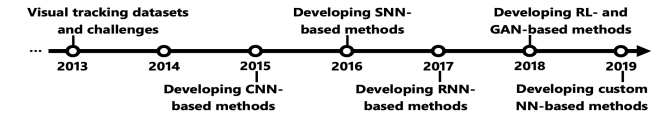
24

Deep Learning for Visual Tracking



25

Deep Learning for Visual Tracking

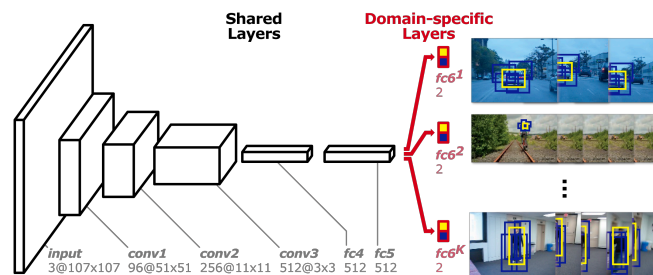


Recent history of Visual tracking

26

Deep Learning for Visual Tracking

CNN based model: MDNet (Multiple Domain)



Learning Multi-Domain Convolutional Neural Networks for Visual Tracking

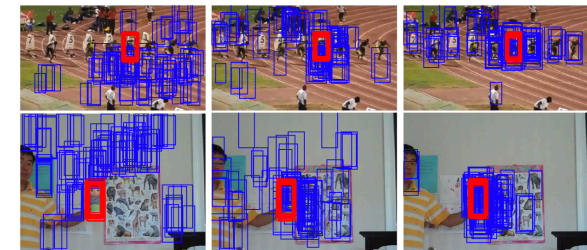
Hyeonseob Nam
Bohyung Han
POSTECH
The Winner of The VOT2015 Challenge

27

Deep Learning for Visual Tracking

CNN based model: MDNet (Multiple Domain)

Selected the domain branch and fine-tuning it according to the target



(a) 1st minibatch (b) 5th minibatch (c) 30th minibatch

Learning Multi-Domain Convolutional Neural Networks for Visual Tracking

Hyeonseob Nam
Bohyung Han
POSTECH
The Winner of The VOT2015 Challenge

28

Deep Learning for Visual Tracking

CNN based model: MDNet (Multiple Domain)

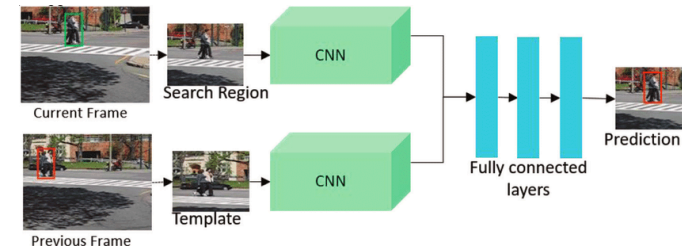


29

Deep Learning for Visual Tracking

SNN based model: GOTURN

Generic Object Tracking Using Regression Networks

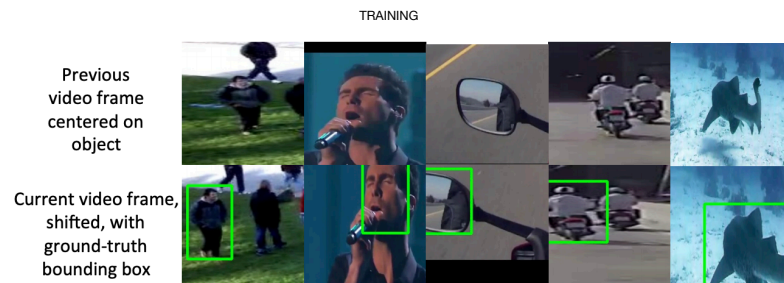


30

Deep Learning for Visual Tracking

SNN based model: GOTURN

Generic Object Tracking Using Regression Networks



Held, David, Sebastian Thrun, et Silvio Savarese. « Learning to Track at 100 FPS with Deep Regression Networks ». CoRR abs/1604.01802 (2016). <http://arxiv.org/abs/1604.01802>.

31

Deep Learning for Visual Tracking

multi-object tracking

Based on Tracking-by-detection

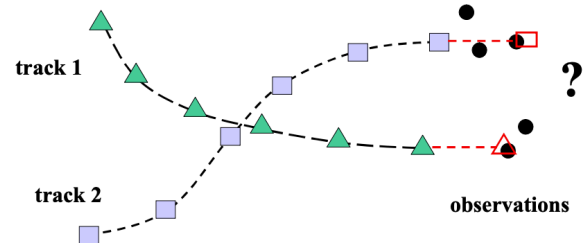
- 1) Object detection
- 2) Metric estimation between detected objects and targets (set of objects with the same identity)
- 3) Association between object and target
- 4) target birth, death and loss.

32

Deep Learning for Visual Tracking

multi-object tracking

Intuition: predict next position along each track.

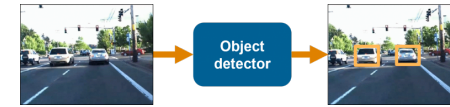


How to determine which observations to add to which track?

33

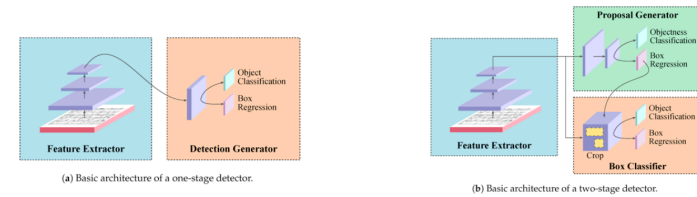
Deep Learning for Visual Tracking

Object detection networks



Two main object detector structures exist:

- One-Stage Detectors
- Two-Stage Detectors



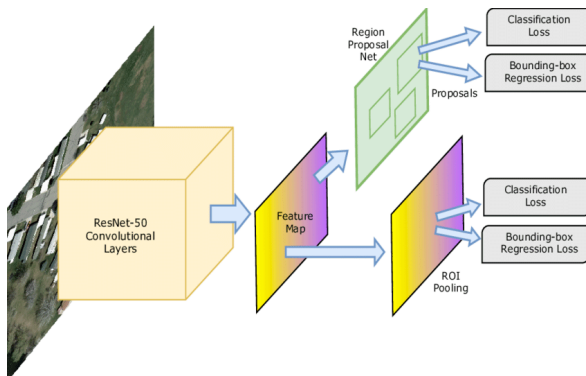
34

Deep Learning for Visual Tracking

Object detection networks

Example of two-stages-detector: Faster-Rcnn

Ren, Shaoqing, Kaiming He, Ross Girshick, et Jian Sun. « Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks ». In *Advances in Neural Information Processing Systems 28*, édité par C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, et R. Garnett, 91-99. Curran Associates, Inc., 2015. <http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf>.



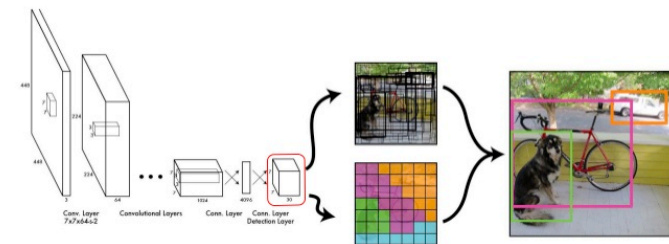
35

Deep Learning for Visual Tracking

Object detection networks

Example of one-stage-detector: YOLO

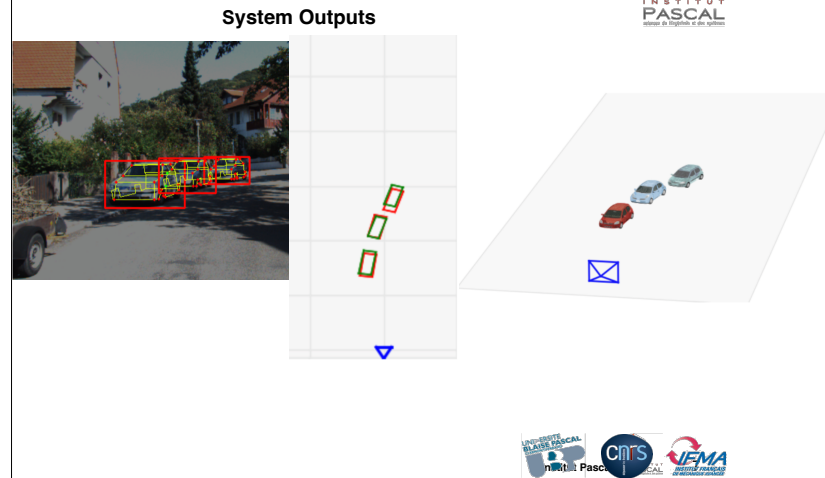
YOLO: You Only Look Once



Redmon, Joseph, Santosh Divvala, Ross Girshick, et Ali Farhadi. « You Only Look Once: Unified, Real-Time Object Detection ». In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

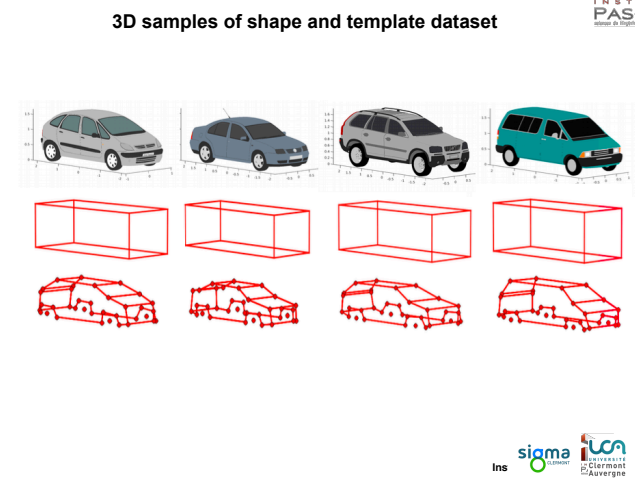
36

Deep Learning for 3D vehicle understanding from monocular images:
toward many-task networks



37

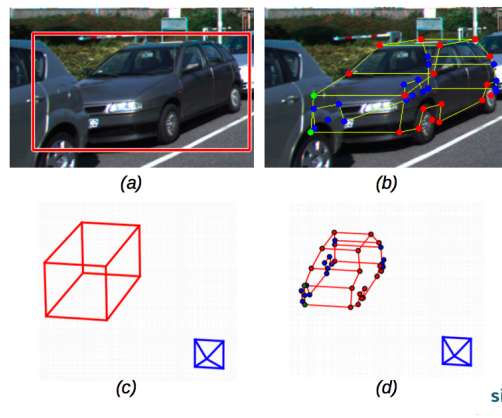
Deep Learning for 3D vehicle understanding from monocular images:
toward many-task networks



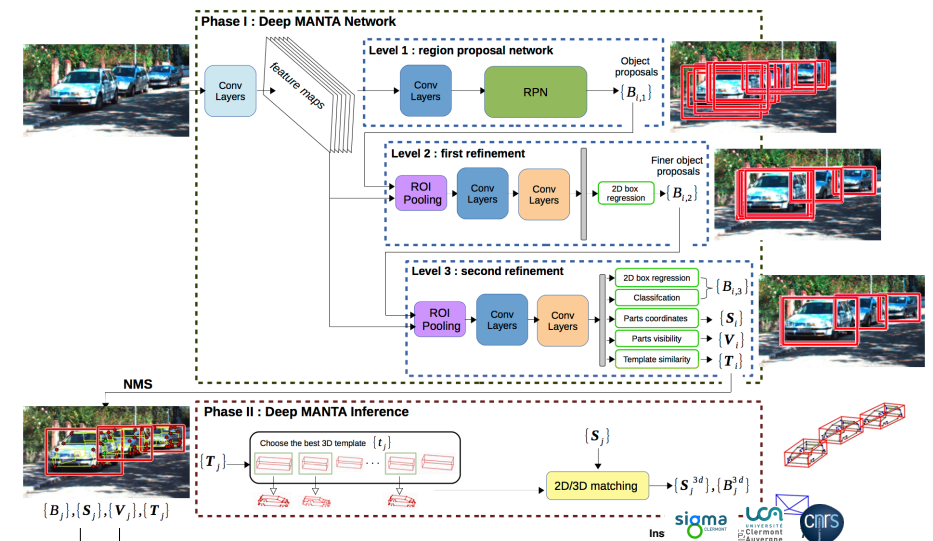
38

Deep Learning for 3D vehicle understanding from monocular images:
toward many-task networks

**Bounding box and part detection
(with visibility estimation, green and blue)**



39



40

Deep Learning for 3D vehicle understanding from monocular images

Loss functions

RPN Loss

Detection loss

Parts Loss

Visibility Loss

Template similarity loss

$$\mathcal{L} = \mathcal{L}^1 + \mathcal{L}^2 + \mathcal{L}^3$$

with

$$\mathcal{L}^1 = \mathcal{L}_{rpn},$$

$$\mathcal{L}^2 = \sum_i \mathcal{L}_{det}^2(i) + \mathcal{L}_{parts}^2(i),$$

$$\mathcal{L}^3 = \sum_i \mathcal{L}_{det}^3(i) + \mathcal{L}_{parts}^3(i) + \mathcal{L}_{vis}(i) + \mathcal{L}_{temp}(i),$$

41

Deep Learning for 3D vehicle understanding from monocular images

Experiments (Kitti Dataset)

42

Deep Learning for Visual Tracking

Object detection networks

2020: Using Transformers for object detection

Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, et Sergey Zagoruyko. *End-to-End Object Detection with Transformers*, 2020.

43

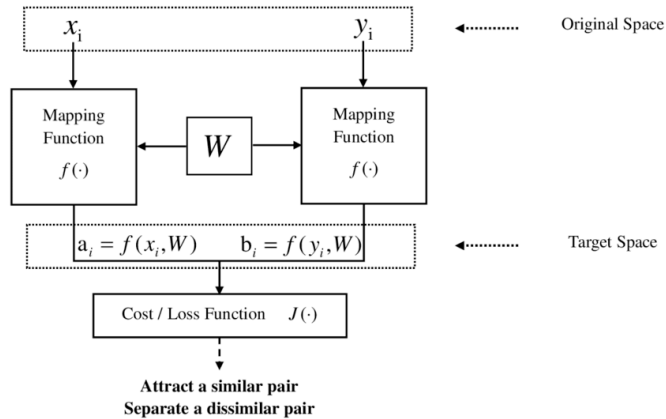
Deep Learning for Visual Tracking

Association: define metric and match objects and targets

44

Deep Learning for Visual Tracking

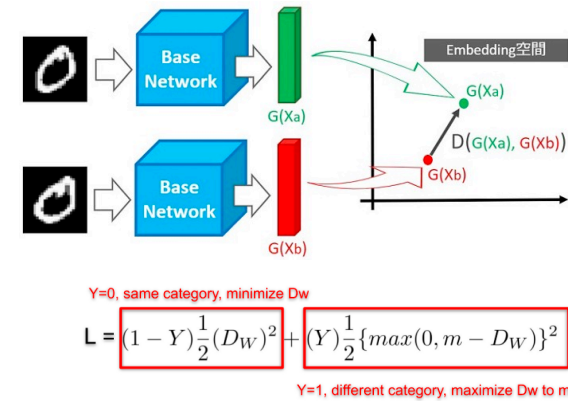
Association: define metric



45

Deep Learning for Visual Tracking

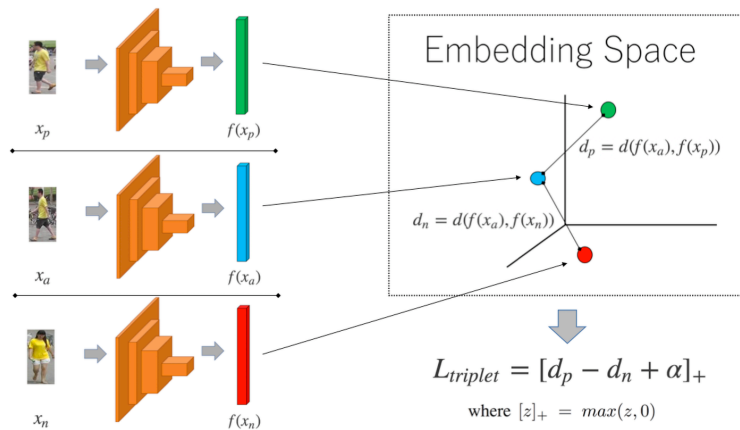
Association: define metric



46

Deep Learning for Visual Tracking

Association:
define metric



47

Deep Learning for Visual Tracking

Association: define metric and match objects and targets (association matrix)

We have N objects in previous frame and M objects in current frame. We can build a table of match scores $m(i, j)$ for $i=1 \dots N$ and $j=1 \dots M$. For now, assume $M=N$.

	1	2	3	4	5
1	0.95	0.76	0.62	0.41	0.06
2	0.23	0.46	0.79	0.94	0.35
3	0.61	0.02	0.92	0.92	0.81
4	0.49	0.82	0.74	0.41	0.01
5	0.89	0.44	0.18	0.89	0.14

problem: choose a 1-1 correspondence that maximizes sum of match scores.

48

Deep Learning for Visual Tracking

Association: define metric and match objects and targets (association matrix)

5x5 matrix of match scores

0.95	0.76	0.62	0.41	0.06
0.23	0.46	0.79	0.94	0.35
0.61	0.02	0.92	0.92	0.81
0.49	0.82	0.74	0.41	0.01
0.89	0.44	0.18	0.89	0.14

working from left to right, choose one number from each column, making sure you don't choose a number from a row that already has a number chosen in it.

How many ways can we do this?

$$5 \times 4 \times 3 \times 2 \times 1 = 120 \text{ (N factorial)}$$

49

Deep Learning for Visual Tracking

Association: define metric and match objects and targets (association matrix)

0.95	0.76	0.62	0.41	0.06
0.23	0.46	0.79	0.94	0.35
0.61	0.02	0.92	0.92	0.81
0.49	0.82	0.74	0.41	0.01
0.89	0.44	0.18	0.89	0.14

score: 2.88

0.95	0.76	0.62	0.41	0.06
0.23	0.46	0.79	0.94	0.35
0.61	0.02	0.92	0.92	0.81
0.49	0.82	0.74	0.41	0.01
0.89	0.44	0.18	0.89	0.14

score: 2.52

0.95	0.76	0.62	0.41	0.06
0.23	0.46	0.79	0.94	0.35
0.61	0.02	0.92	0.92	0.81
0.49	0.82	0.74	0.41	0.01
0.89	0.44	0.18	0.89	0.14

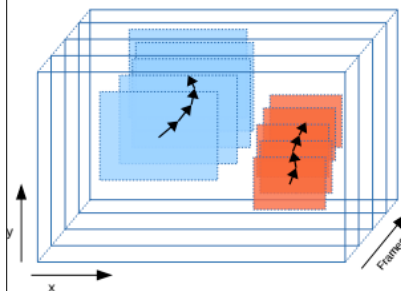
score: 4.14

50

Deep Learning for Visual Tracking

Object detection networks

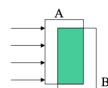
SORT: Tracking-by-detection



State Vector : $\mathbf{x} = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]^T$,
Position, scale, ratio

Trajectory prediction: Kalman filter

Association: IOU distance and Hungarian Algorithm



$$\text{score} = \frac{2 * \text{area}(A \text{ and } B)}{\text{area}(A) + \text{area}(B)}$$

Detections			
a1	a2	a3	a4
b1	b2	b3	b4
c1	c2	c3	c4
d1	d2	d3	d4

Bewley, Alex, ZongYuan Ge, Lionel Ott, Fabio Ramos, et Ben Uppert. « Simple Online and Realtime Tracking ». *CoRR* abs/1602.00763 (2016). <http://arxiv.org/abs/1602.00763>.

51

Deep Learning for Visual Tracking

SORT: Tracking-by-detection



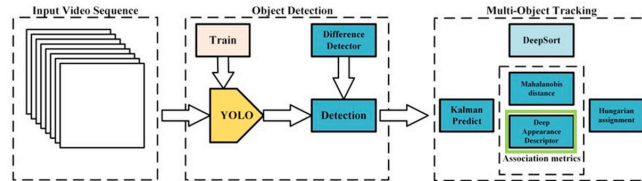
<https://medium.com/neuromation-blog/tracking-cows-with-mask-r-cnn-and-sort-fcd4ad68ec4f>

52

Deep Learning for Visual Tracking

DeepSORT: Tracking-by-detection

SORT WITH DEEP ASSOCIATION METRIC

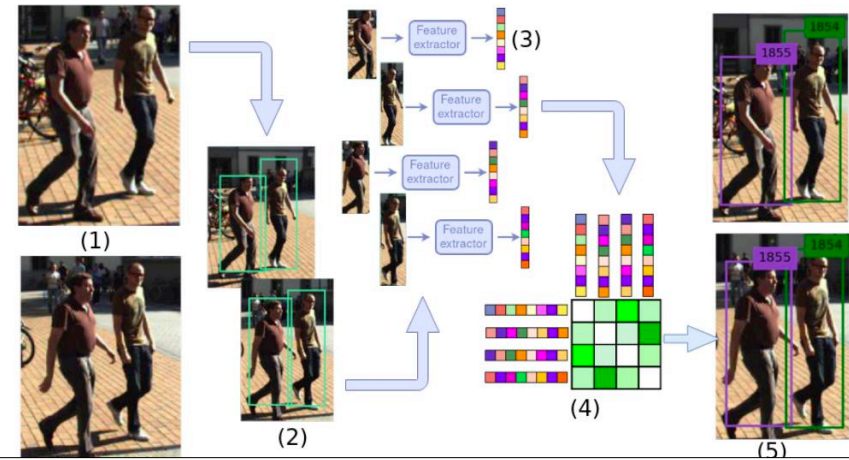


<https://medium.com/neuromation-blog/tracking-cows-with-mask-r-cnn-and-sort-fcd4ad68ec4f>

53

Deep Learning for Visual Tracking

DeepSORT: Tracking-by-detection SORT WITH DEEP ASSOCIATION METRIC



54

Deep Learning for Visual Tracking

DeepSORT: Tracking-by-detection

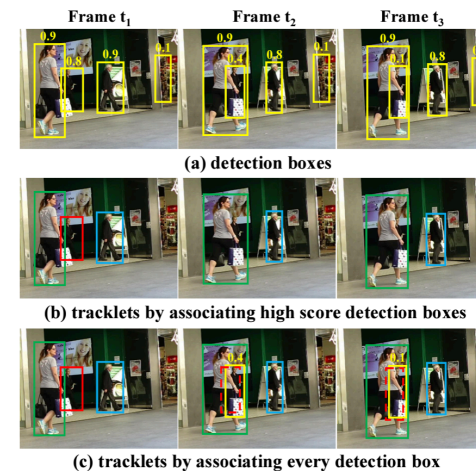
SORT WITH DEEP ASSOCIATION METRIC



55

Deep Learning for Visual Tracking

ByteTrack: Multi Object Tracking



56

Deep Learning for Visual Tracking

Yellow: high score boxes
red: low score boxes

ByteTrack: Multi Object Tracking



57

Deep Learning for Visual Tracking

Yellow: high score boxes
red: low score boxes

ByteTrack: Multi Object Tracking



58

The power of video interlacing

Introduction

The classical tracking-by-detection scheme:

- object detection (for each frame of the video sequence)
- **association (spatio-temporal and/or appearance models)**
- birth and death trajectories algorithms

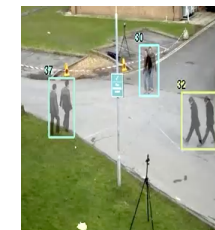
59

The power of video interlacing

The key idea

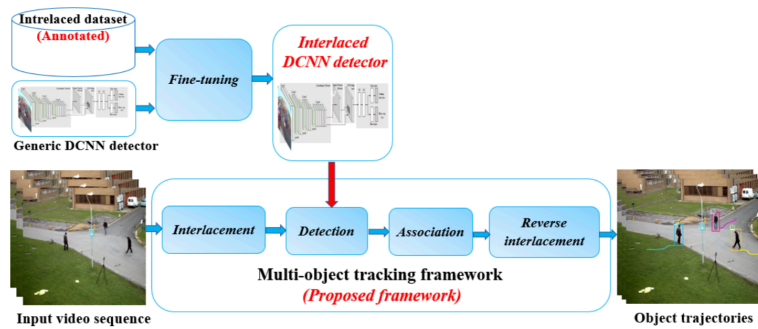
build an **interlaced** video
and train an interlaced
pedestrian detector

- to:
- increase overlapping
between successive
frames
 - learn appearance
association within a
deep convolution
neural network



60

The power of video interlacing



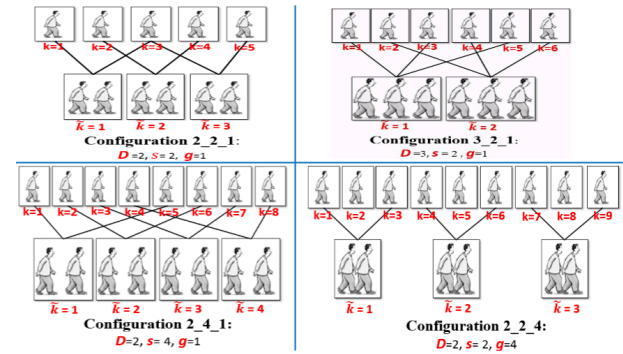
Institut Pascal
PASCAL

61

The power of video interlacing

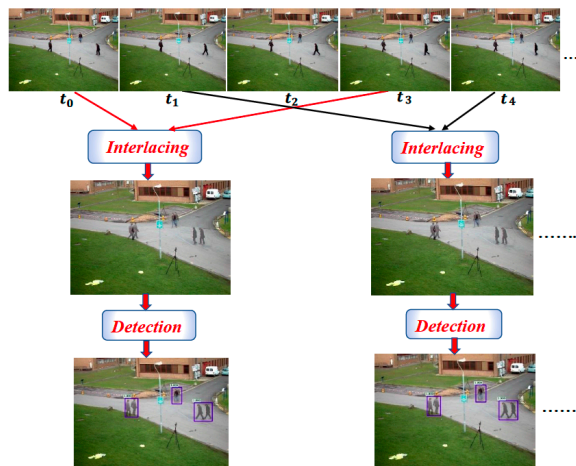
Build a interlaced video

$$\tilde{I}_k(x, y) \doteq \sum_{d=0, \dots, (D-1)} I_{(\tilde{k}g+ds)}(x, y) \cdot \delta(y[D] - d)$$



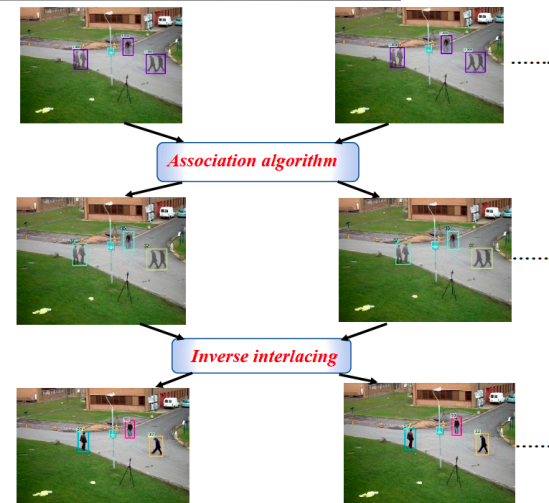
62

The power of video interlacing



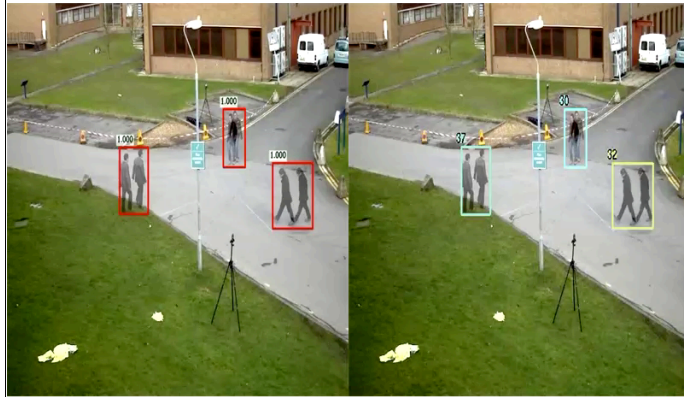
63

The power of video interlacing



64

The power of video interlacing



Detection

Tracking

Institut Pascal

65



66