Learning by tracking: Siamese CNN for robust tracking association

Laura Leal-Taixé[1] Cristian Canton-Ferrer[2] Konrad Schindler[3]

[1] Computer Vision Group, Technical University Munich
[2] Microsoft, Redmond
[3] Photogrammetry and Remote Sensing, ETH Zürich

Motivation

Problem: multiple object tracking. **Framework**: tracking-by-detection

Previous works

Data term

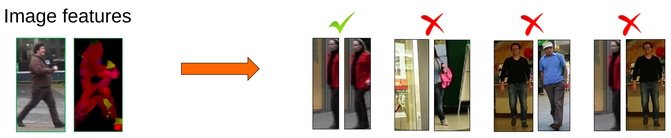
- Distance information
- Appearance information
- Motion models
- Interaction between pedestrians

Complex optimization models

- Conditional Random Fields
- Maximum cliques
- Dual decomposition
- Continuous energy minimization

- ✗ Sequence dependent
- ✗ Hard to balance the terms
- ✗ Hard to optimize
- ✗ Computationally expensive

Can we directly learn the data association?

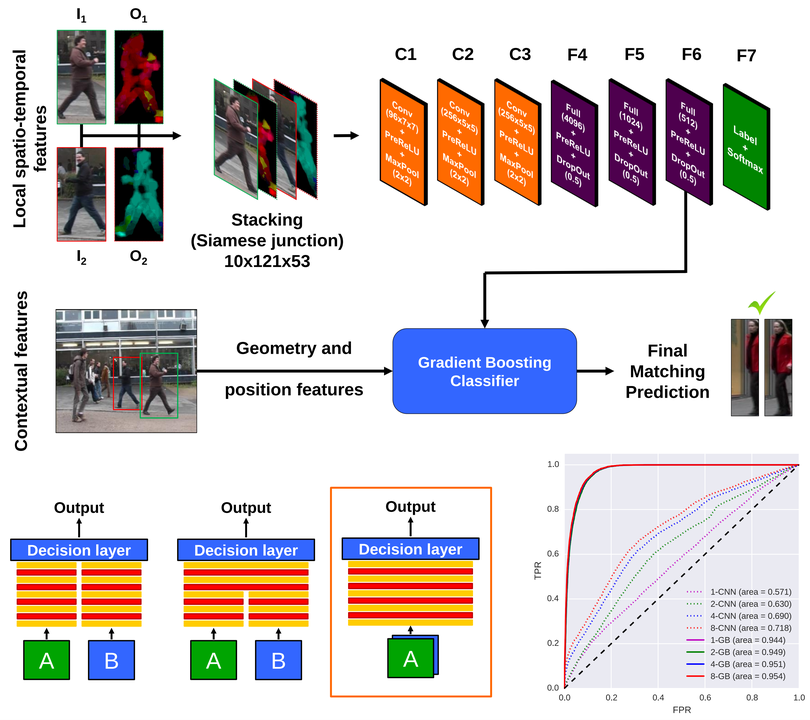


- ✓ No tuning different terms
- ✓ We can use an easy-to-optimize tracking framework
- ✓ Can handle a large variety of sequences

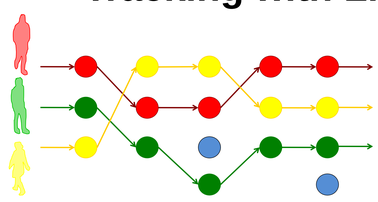
CNN have shown great results for patch comparison

Siamese CNN for patch comparison

Much **richer representation** of the **physical interaction** between pedestrians wrt. hand-designed terms of the Social Force Model.



Tracking with Linear Programming



$$\mathcal{T}* = \underset{\mathcal{T}}{\operatorname{argmin}} \sum_i C_{\text{in}}(i) f_{\text{in}}(i) + \sum_i C_{\text{out}}(i) f_{\text{out}}(i) + \sum_i C_{\text{det}}(i) f(i) + \sum_{i,j} C_{\text{t}}(i,j) f(i,j)$$

Find the set of trajectories by solving the minimum cost-flow problem [1]

✓ Easy to optimize: k-shortest paths, Simplex method.

Results

Quantitative evaluation on MOTChallenge 2015 test set [2], *www.motchallenge.net*

Method	TA	TP	MT	ML	IDsw	FP
NOMT	**33.7**	**71.9**	12.2	44.0	442	7762
MHT-DAM	32.4	71.8	**16.0**	43.8	435	9064
MDP	30.3	71.3	13.0	**38.4**	680	9717
SiameseCNN	29.0	71.2	8.5	48.4	639	**5160**
LP-SSVM	25.2	71.7	5.8	53.0	849	8369
ELP	25.0	71.2	7.5	43.8	1396	7345
JPDA-m	23.8	68.2	5.0	58.1	**365**	6373
MotiCon	23.1	70.9	4.7	52.0	1018	10404
SegTrack	22.5	71.7	5.8	63.9	697	7890
LP2D	19.8	71.2	6.7	41.2	1649	11580
DCO-X	19.6	71.4	5.1	54.9	521	10652
CEM	19.3	70.7	8.5	46.5	813	14180
RMOT	18.6	69.6	5.3	53.3	684	12473
SMOT	18.2	71.2	2.8	54.8	1148	8780
ALExTRAC	17.0	71.2	3.9	52.4	1859	9233
TBD	15.9	70.9	6.4	47.9	1939	14943
TC-ODAL	15.1	70.5	3.2	55.8	637	12970
DP-NMS	14.5	70.8	6.0	40.8	4537	13171
LDCT	4.7	71.7	11.4	32.5	12348	14066

Same Linear Programming framework!

- LP2D: uses only distance between bounding boxes
- MotiCon: learned motion context from manually designed image features [3]
- SiameseCNN: proposed learned costs

Conclusions

- We proposed a way to **estimate detection associations** in the context of multiple target tracking.
- We **stacked the image and optical flow** channels and **fused contextual features** with the last fully connected layer for increased accuracy.
- We showed that a simple tracking framework outperforms complex models when **fed with accurate information**.

[1] L. Leal-Taixé, G. Pons-Moll, B. Rosenhahn. *Everybody needs somebody: modeling social and grouping behavior on a linear programming multiple people tracker.* ICCVW 2011.

[2] L. Leal-Taixé, A. Milan, I. Reid, S. Roth and K. Schindler. *MOTChallenge 2015: Towards a benchmark for multi-target tracking*. arXiv:1504.01942

[3] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn and S. Savarese. *Learning an image-based motion context for multiple people tracking*. CVPR 2014.