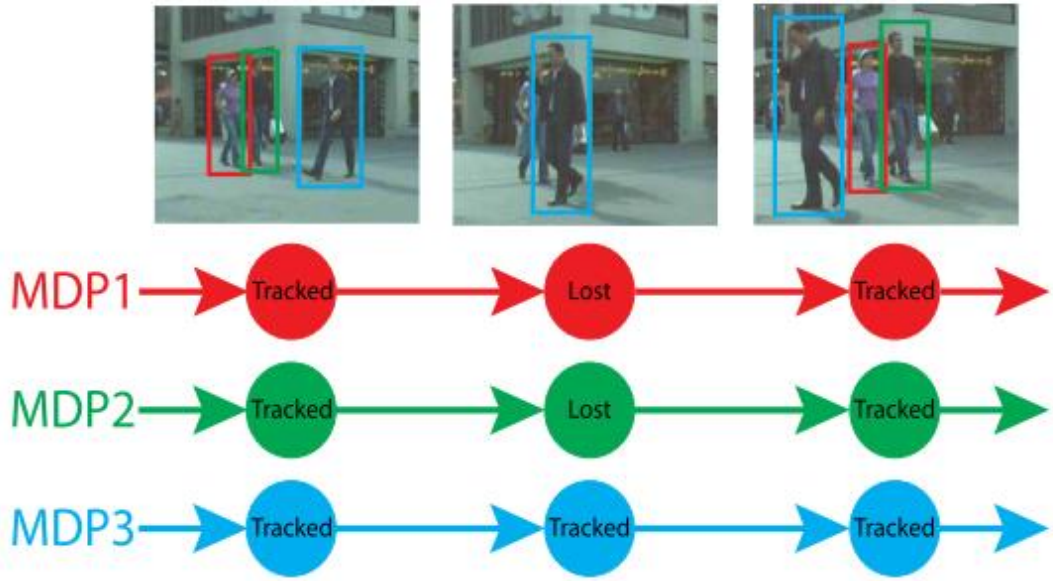
Learning to Track: Online Multi-object Tracking by Decision Making

Yu Xiang, Alexandre Alahi and Silvio Savarese

ICCV 2015

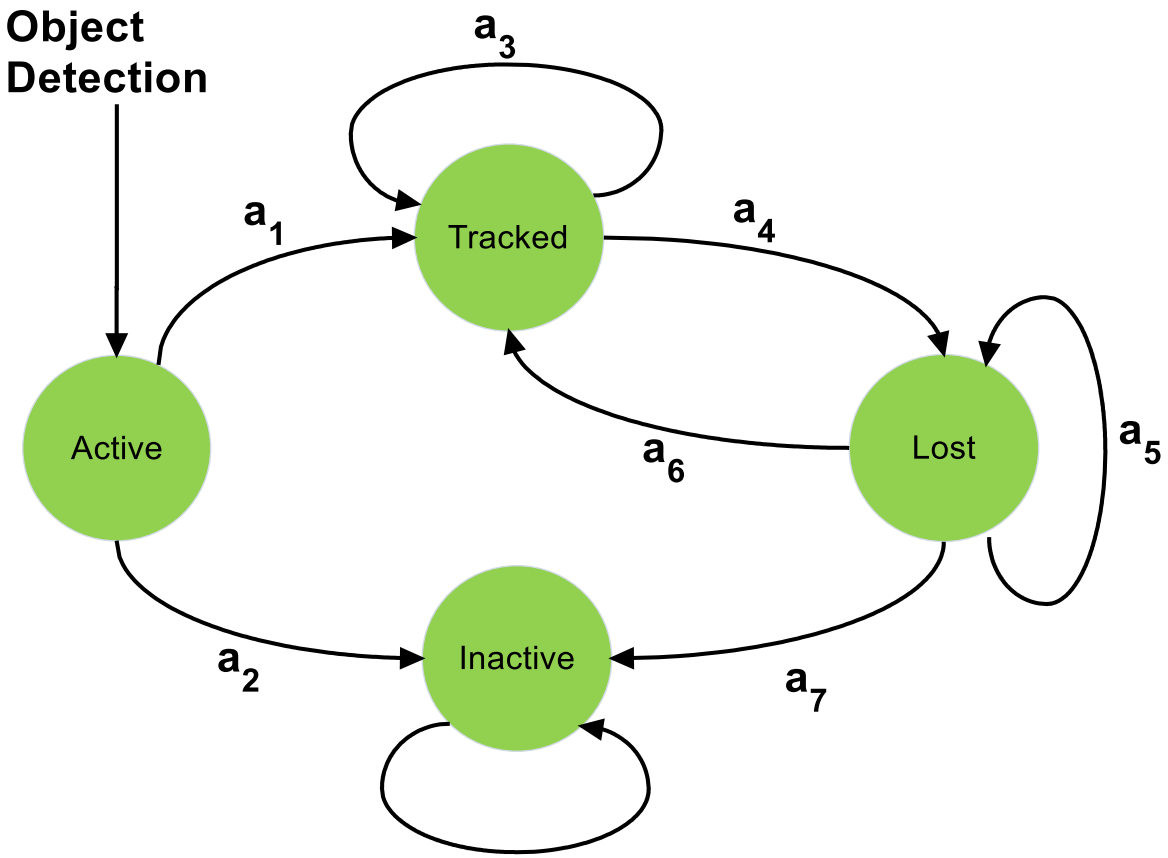
Introduction

- Online Multi Object Tracking
- Detector and tracker working in parallel
- Target lifetime modeled as a Markov Decision Process (MDP)
- Reinforcement Learning used to learn similarity function for data association



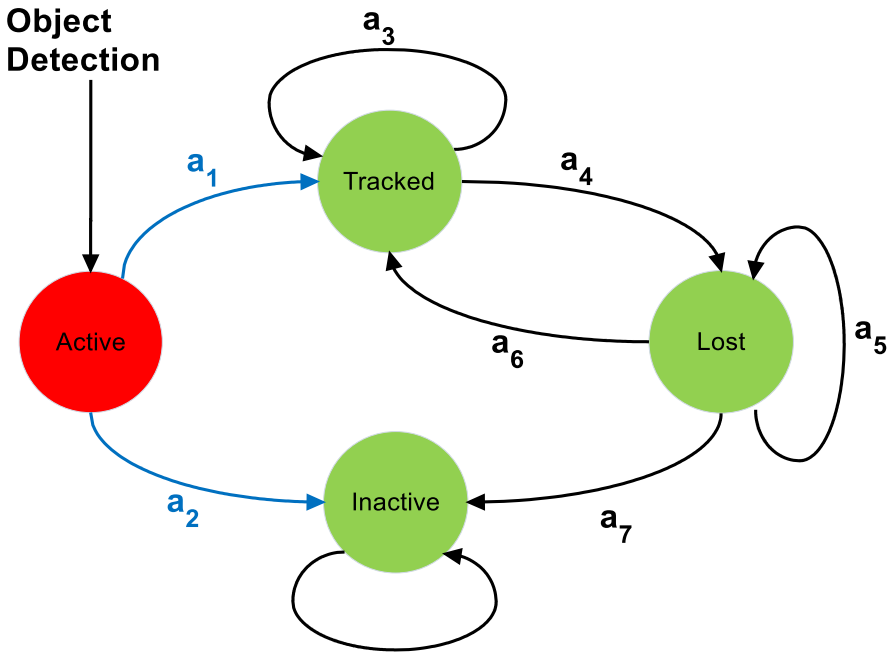
Markov Decision Process

- Four states – active, tracked, lost, inactive
- Seven actions
- Deterministic Transition Function



Active State Policy

- Initial state when an object is first detected
- Transition to **tracked** (a_1) or **inactive** (a_2)

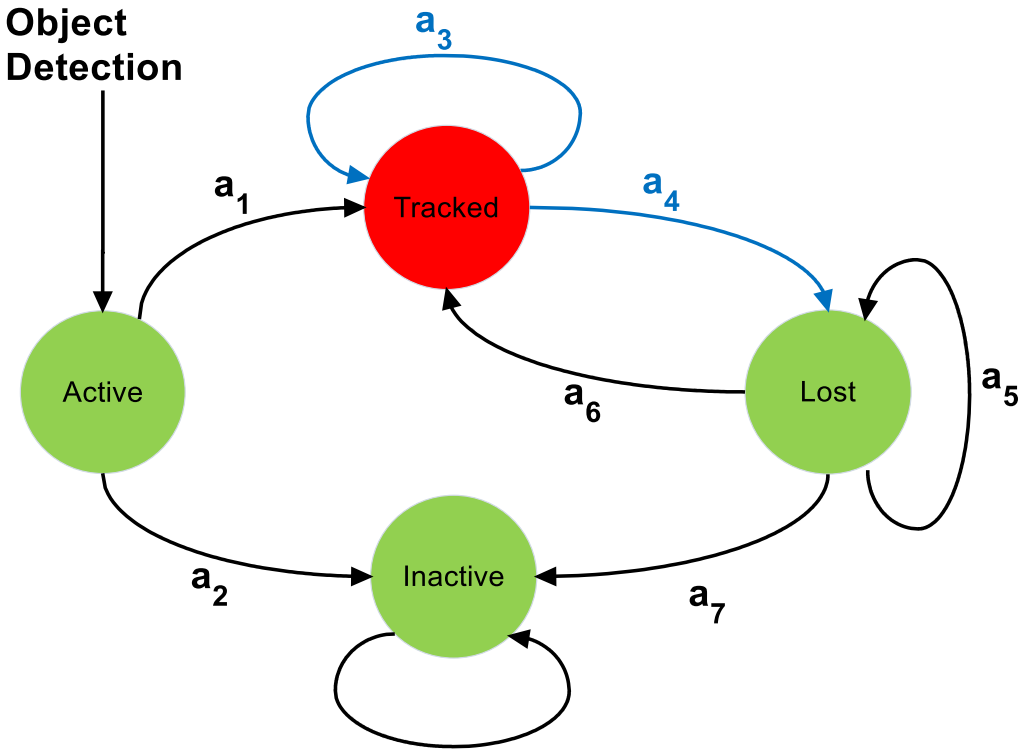


- Train binary **SVM** using a 5D feature vector:
  - x, y coordinates, width, height, detection score
- Equivalent to learning reward function:

$$R_{\text{Active}}(s,a) = y(a)\left(\mathbf{w}_{\text{Active}}^{T}\phi_{\text{Active}}(s) + b_{\text{Active}}\right)$$

Tracked State Policy

- An existing object that is visible in the current frame
- Remain **tracked** (a_3) or transition to **lost** (a_4)



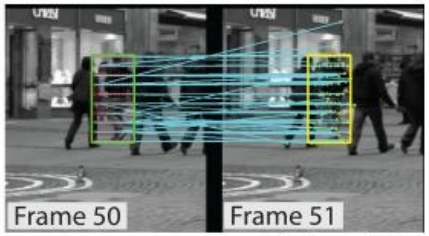
- Decision depends on the tracking method used
  - **TLD**[Kalal12] used here

Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-Learning-Detection," TPAMI 2012

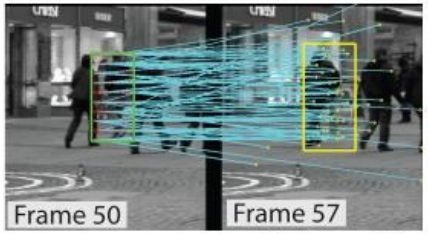
Tracked State Policy – Tracking

- Template represented by image patch under the bounding box
- **Optical flow** of densely uniformly sampled points in template
  - Iterative Lucas Kanade with Pyramids
- Forward - Backward (**FB**) estimation used to estimate stability

(a) target template



(b) stable prediction



(c) unstable prediction

- Patches from all tracked frames are collected
  - Used as target **history** for data association
- **Lazy** update
  - Template updated only when target gets lost

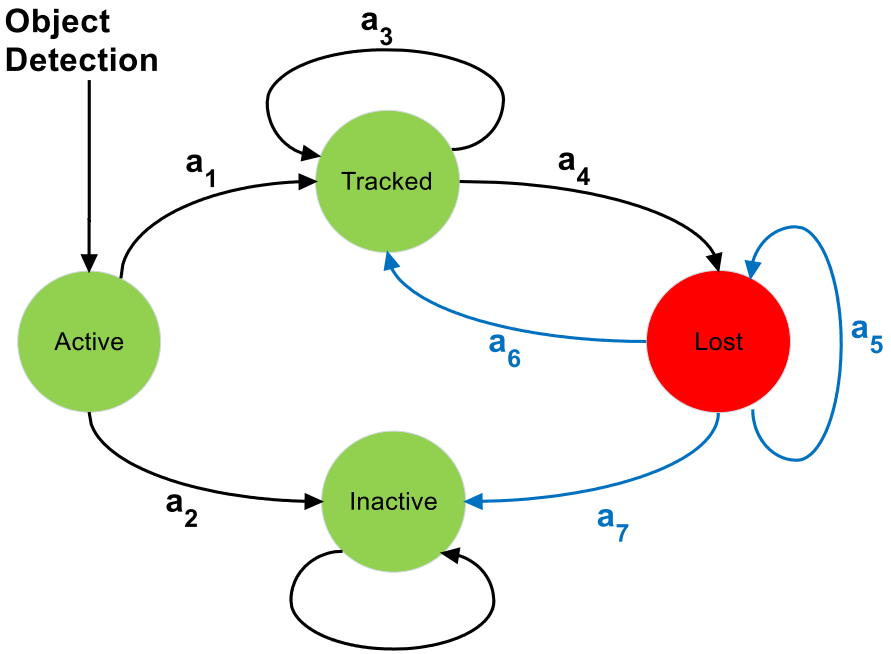
Tracked State Policy – Decision Making

- 2D feature vector used for decision making
- Median FB Error
  - Euclidean distance between initial point and FB prediction
- Mean bounding box overlap between the target history and corresponding detections
  - Optical flow can continue tracking false detections
  - False objects cannot be detected consistently
- Equivalent reward function:

$$R_{\text{Tracked}}(s,a) = \begin{cases} y(a), \text{ if } e_{\text{medFB}} < e_0 \text{ and } o_{\text{mean}} > o_0 \\ -y(a), \text{ otherwise,} \end{cases}$$

Lost State Policy

- Target is occluded or goes out of view of the camera
- Remain **lost** (a_5) or transition to **tracked** (a_6) or **inactive** (a_7)



- Inactive if lost for more than T_{Lost} frames
- Data association used for deciding between lost and tracked

Lost State Policy – Data Association

- Soft margin binary SVM classifier used to generate similarity between target t and detection d

$$f(t,d) = \mathbf{w}^T \phi(t,d) + b$$

- SVM is trained using RL
  - Use existing classifier to track objects on training sequences
  - Update whenever it makes a mistake in data association
- Two types of mistakes:
  - Target is incorrectly associated to a detection - add as **negative** training example
  - Target is not associated with any detection but is visible and detected correctly – add as **positive** training example
- Equivalent reward function:

$$R_{\text{Lost}}(s,a) = y(a)\left(\max_{k=1}^{M}\left(\mathbf{w}^T \phi(t,d_k) + b\right)\right)$$

Lost State Policy – Features

Type	Notation	Feature Description
FB error	$\phi_1, \cdots, \phi_5$	Mean of the median forward-backward errors from the entire, left half, right half, upper half and lower half of the templates in optical flow
NCC	ϕ_6	Mean of the median Normalized Correlation Coefficients (NCC) between image patches around the matched points in optical flow
	ϕ_7	Mean of the NCC between image patches of the detection and the predicted bounding boxes from optical flow
Height ratio	ϕ_8	Mean of the ratios in bounding box height between the detection and the predicted bounding boxes from optical flow
	ϕ_9	Ratio in bounding box height between the target and the detection
Overlap	ϕ_{10}	Mean of the bounding box overlaps between the detection and the predicted bounding boxes from optical flow
Score	ϕ_{11}	Normalized detection score
Distance	ϕ_{12}	Euclidean distance between the centers of the target and the detection after motion prediction of the target with a linear velocity model

Lost State Policy – RL Algorithm

input : Video sequences $\mathcal{V} = \{v_i\}_{i=1}^{N}$, ground truth trajectories $\mathcal{T}_i = \{t_{ij}\}_{j=1}^{N_i}$ and object detection $\mathcal{D}_i = \{d_{ij}\}_{j=1}^{N_i'}$ for video $v_i, i = 1, \ldots, N$

output: Binary classifier $(\mathbf{w}, b)$ for data association

Initialization: $\mathbf{w} \leftarrow \mathbf{w}_0, b \leftarrow b_0, \mathcal{S} \leftarrow \emptyset$
repeat
  foreach *video* v_i *in* $\mathcal{V}$ **do**
    foreach *target* t_{ij} *in* v_i **do**
      Initialize the MDP in Active ;
      $l \leftarrow$ index of the 1st frame t_{ij} correctly detected ;
      Transfer the MDP to Tracked, and initial the target template ;
      while $l \leq$ *index of last frame of* t_{ij} **do**
        Follow the current policy and choose an action a ;
        Compute the action a_{gt} indicated by the ground truth ;
        if *Current state is Lost and* $a \neq a_{gt}$ **then**
          Decide the label y_k of the pair (t_{ij}^{l}, d_k) ;
          $\mathcal{S} \leftarrow \mathcal{S} \cup \{(\phi(t_{ij}^{l}, d_k), y_k)\}$;
          $(\mathbf{w}, b) \leftarrow$ solution of Eq. (4) on $\mathcal{S}$;
          break ;
        else
          Execute action a ;
          $l \leftarrow l + 1$;
        end
      end
      if $l >$ *index of last frame of* t_{ij} **then**
        Mark target t_{ij} as successfully tracked;
      end
    end
  end
until *all targets are successfully tracked*;

Update Classifier by solving soft margin optimization problem:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{M}\xi_k$$

$$\text{s.t. } y_k\left(\mathbf{w}^T\phi(t_k, d_k) + b\right) \geq 1 - \xi_k, \xi_k \geq 0, \forall k$$

Overall MDP Tracking Algorithm

input : A video sequence v and object detection $\mathcal{D} = \{d_k\}_{k=1}^{N}$ for v,
binary classifier $(\mathbf{w}, b)$ for data association
output: Trajectories of targets $\mathcal{T} = \{t_i\}_{i=1}^{M}$ in the video

Initialization: $\mathcal{T} \leftarrow \emptyset$;
foreach *video frame* l *in* v **do**
    // process targets in tracked states
    foreach *tracked target* t_i *in* $\mathcal{T}$ **do**
        Follow the policy, move the MDP of t_i to the next state ;
    end
    // process targets in lost states
    foreach *lost target* t_i *in* $\mathcal{T}$ **do**
        foreach *detection* d_k *not covered by any tracked target* **do**
            Compute $f(t_i, d_k) = \mathbf{w}^T \phi(t_i, d_k) + b$;
        end
    end
    Data association with Hungarian algorithm for the lost targets ;
    foreach *lost target* t_i *in* $\mathcal{T}$ **do**
        Follow the assignment, move the MDP of t_i to the next state ;
    end
    // initialize new targets
    foreach *detection* d_k *not covered by any tracked target in* $\mathcal{T}$ **do**
        Initialize a MDP for a new target t with detection d_k ;
        if *action* a_1 *is taken following the policy* **then**
            Transfer t to the tracked state ;
            $\mathcal{T} \leftarrow \mathcal{T} \cup \{t\}$;
        else
            Transfer t to the inactive state ;
        end
    end
end

Datasets

- Multi Object Tracking (MOT) benchmark
- 11 training and 11 testing sequences
- Test sequence ground truth not available
  - 6 of the training sequences used for validation
- Training sequence captured in similar scenario

Training	Testing
Validation on MOT Benchmark	
TUD-Stadtmitte	TUD-Campus
ETH-Bahnhof	ETH-Sunnyday, ETH-Pedcross2
ADL-Rundle-6	ADL-Rundle-8, Venice-2
KITTI-13	KITTI-17
Testing on MOT Benchmark	
TUD-Stadtmitte, TUD-Campus	TUD-Crossing
PETS09-S2L1	PETS09-S2L2, AVG-TownCentre
ETH-Bahnhof, ETH-Sunnyday, ETH-Pedcross2	ETH-Jelmoli, ETH-Linthescher, ETH-Crossing
ADL-Rundle-6, ADL-Rundle-8	ADL-Rundle-1, ADL-Rundle-3
KITTI-13, KITTI-17	KITTI-16, KITTI-19
Venice-2	Venice-1

Evaluation Metrics

Metric	Description
MOTA	**Multiple Object Tracking Accuracy** - combines three error sources: false positives, missed targets and identity switches
MOTP	**Multiple Object Tracking Precision** - misalignment between annotated and predicted bounding boxes
MT	**Mostly tracked targets** - percentage of ground truth trajectories that are covered by tracking output for at least 80% of their respective life span
ML	**Mostly lost targets** - Percentage of ground truth trajectories that are covered by tracking output less than 20% of their respective life span
FP	Total number of **false positives**
FN	Total number of **false negatives** (missed targets)
IDS	Total number of **identity switches**
Frag	Total number of times a trajectory is **fragmented** (i.e. interrupted during tracking)
Hz	Number of frames processed in one second

Higher is better

Lower is better

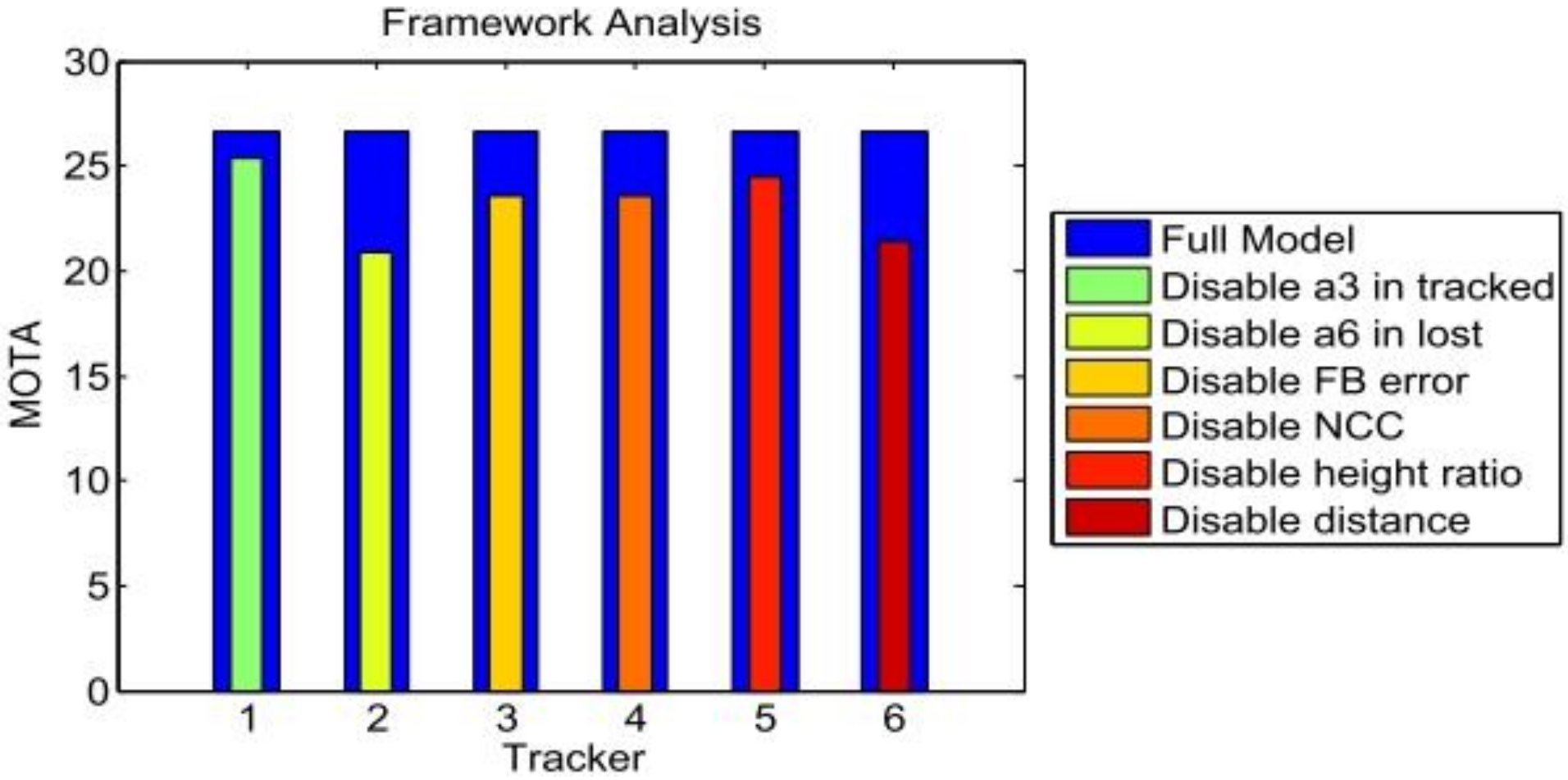
Analysis on Validation Set

- Number of templates in the history (K)

K	MOTA	MOTP	MT	ML	FP	FN	IDS	Frag
1	24.7	73.2	10.3	55.1	3,597	13,651	147	303
2	25.7	73.5	9.8	53.4	3,548	13,485	121	349
3	23	73.6	8.5	56	3,727	13,907	134	325
4	26.3	**73.9**	9.8	53.8	3,191	13,726	**91**	300
5	**26.7**	73.7	**12**	53	3,386	**13,415**	111	331
6	19.5	73.7	5.6	68.8	3,393	14,920	269	321
7	26.1	73.6	10.7	55.6	3,092	13,838	132	306
8	25.8	73.8	10.7	55.6	3,221	13,785	122	305
9	**26.7**	73.6	**12**	**51.7**	3,290	13,491	133	328
10	26.6	73.8	9.8	55.1	**2,691**	14,130	123	**276**
11	25.3	73.5	**12**	52.1	3,672	13,436	136	317
12	24.8	73.4	11.5	55.6	3,637	13,585	139	321

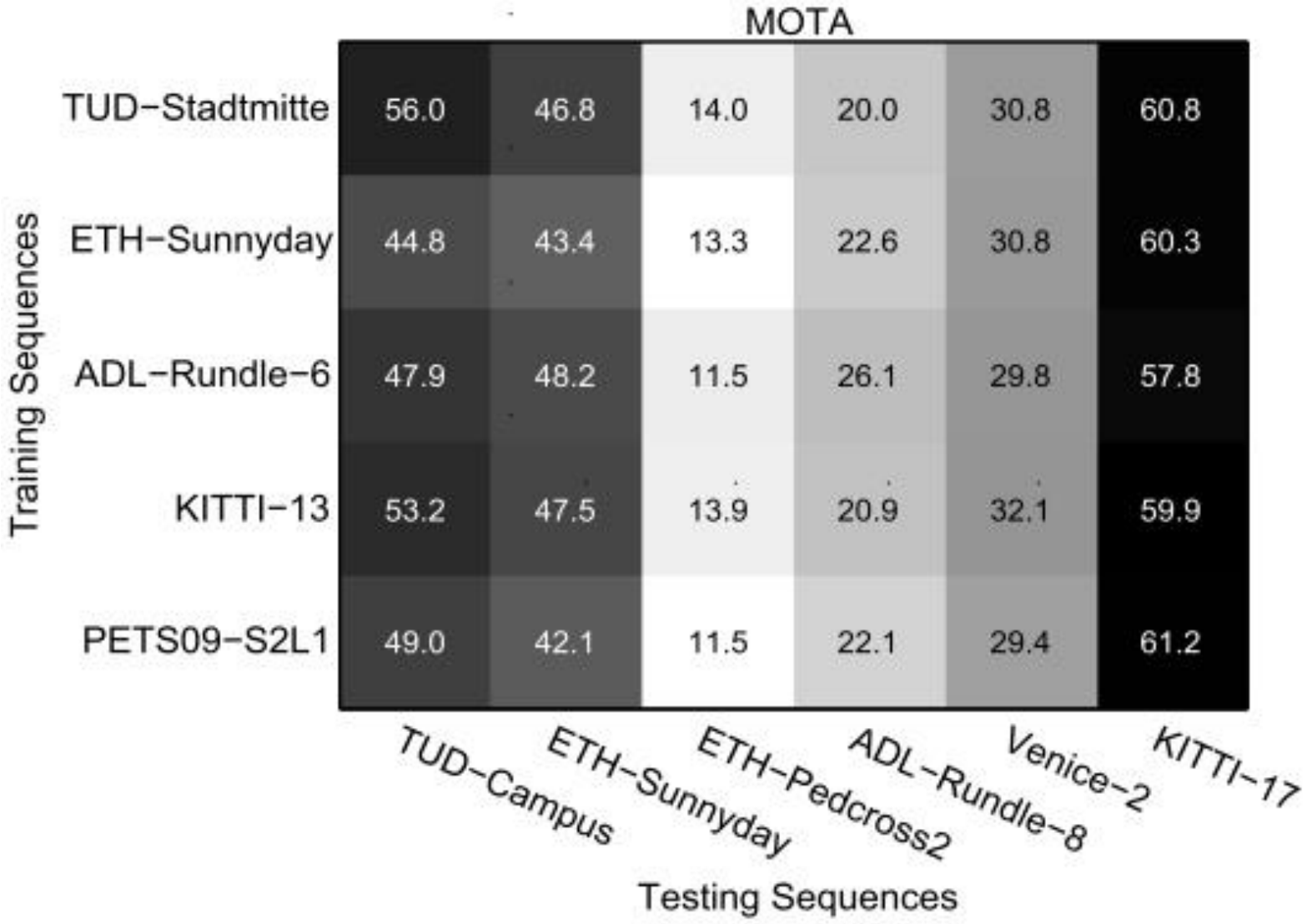
Analysis on Validation Set

- Disabling different components



Analysis on Validation Set

- Cross domain training and testing with MOTA



Tested Trackers

- **DP NMS**: H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, 'Globally-optimal greedy algorithms for tracking a variable number of objects', CVPR 2011
- **TC ODAL**: S.-H. Bae and K.-J. Yoon, 'Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning', CVPR 2014
- **TBD**: A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, '3d traffic scene understanding from movable platforms', TPAMI 2014
- **SMOT**: C. Dicle, O. I. Camps, and M. Sznaier, 'The way they move: Tracking multiple targets with similar appearance', ICCV 2013
- **RMOT**: J. H. Yoon, M.-H. Yang, J. Lim, and K.-J. Yoon, 'Bayesian multi-object tracking using motion context from multiple objects', WACV 2015
- **CEM**: A. Milan, S. Roth, and K. Schindler, 'Continuous energy minimization for multitarget tracking', TPAMI 2014
- **SegTrack**: A. Milan, L. Leal-Taix´e, K. Schindler, and I. Reid, 'Joint tracking and segmentation of multiple targets', CVPR 2015
- **MotiCon**: L. Leal-Taix´e, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese, 'Learning an image-based motion context for multiple people tracking', CVPR 2014

Results

- **MDP REL**: MDP Reinforcement Learning
- **MDP OFL**: MDP Offline Learning
  - Link detections using ground truth to form target trajectories
  - Pairs of target and detection that should/should not be linked between adjacent frames

Tracker	Tracking Mode	Learning Mode	MOTA	MOTP	MT	ML
DP NMS	Batch	N/A	14.5	70.8	6.00%	40.80%
TC ODAL	Online	Online	15.1	70.5	3.20%	55.80%
TBD	Batch	Offline	15.9	70.9	6.40%	47.90%
SMOT	Batch	N/A	18.2	71.2	2.80%	54.80%
RMOT	Online	N/A	18.6	69.6	5.30%	53.30%
CEM	Batch	N/A	19.3	70.7	8.50%	46.50%
SegTrack	Batch	Offline	22.5	**71.7**	5.80%	63.90%
MotiCon	Batch	Offline	23.1	70.9	4.70%	52.00%
MDP OFL	Online	Offline	30.1	71.6	10.40%	41.30%
MDP REL	Online	Online	**30.3**	71.3	**13.00%**	**38.40%**

Results

- **MDP REL**: MDP Reinforcement Learning
- **MDP OFL**: MDP Offline Learning
  - Link detections using ground truth to form target trajectories
  - Pairs of target and detection that should/should not be linked between adjacent frames

Tracker	Tracking Mode	FP	FN	IDS	Frag	Hz
DP NMS	Batch	13,171	34,814	4,537	3,090	**444.8**
TC ODAL	Online	12,970	38,538	**637**	1,716	1.7
TBD	Batch	14,943	34,777	1,939	1,963	0.7
SMOT	Batch	8,780	40,310	1,148	2,132	2.7
RMOT	Online	12,473	36,835	684	1,282	7.9
CEM	Batch	14,180	34,591	813	1,023	1.1
SegTrack	Batch	**7,890**	39,020	697	**737**	0.2
MotiCon	Batch	10,404	35,844	1,018	1,061	1.4
MDP OFL	Online	8,789	33,479	690	1,301	0.8
MDP REL	Online	9,717	**32,422**	680	1,500	1.1

Results


TUD-Crossing #31


PETS09-S2L2 #68


PETS09-S2L2 #111


ETH-Jelmoli #82


ETH-Linthescher #51


ETH-Crossing #97


AVG-TownCentre #52


ADL-Rundle-1 #232


ADL-Rundle-3 #183


Venice-1 #235


KITTI-16 #90, KITTI-19 #281

Thanks !