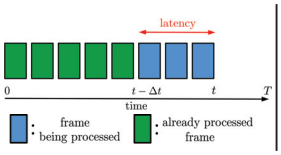
Improving Multi-frame Data Association with Sparse Representations for Robust Near-online Multi-object Tracking

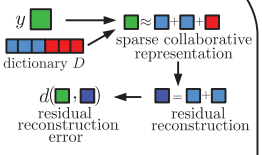
Loïc Fagot-Bouquet[1], Romaric Audigier[1], Yoann Dhome[1], Frédéric Lerasle[2,3]
[1]CEA, LIST, Vision and Content Engineering Laboratory, Point Courrier 173, F-91191 Gif-sur-Yvette, France
[2]CNRS, LAAS, 7, Avenue du Colonel Roche, F-31400 Toulouse, France
[3]Univ. de Toulouse, UPS, LAAS, F-31400 Toulouse, France

❶ Context

Multi-object Tracking by detection



Sliding window-based approach

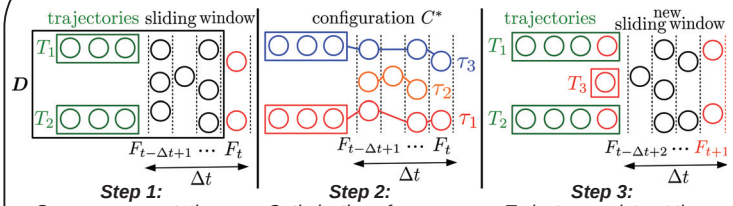


Classification with collaborative representations

Motivations:
- Reasoning on several frames improves performances with only a slight latency.
- Complex appearance models help distinguishing targets.
- **Sparse representations** widely used in **Single Object Tracking** and a few **online Multiple Object Tracking** approaches.

Main contributions:
- Energy based on **sparse representations** for the multi-frame data association. ❸
- Penalty ($l_{\infty,1}$ **norm)** more suited to the case of Multi-object Tracking. ❹

❷ Framework overview

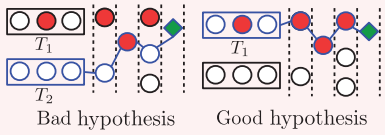


$$E(C) = \theta_{Ob} Ob(C) + \theta_{App} App(C) + \theta_{Mot} Mot(C) + \theta_{Int} Int(C)$$

- $\theta_{Ob} Ob(C)$: Favors configurations involving the most confident detections.
- $\theta_{App} App(C)$: See ❸.
- $\theta_{Mot} Mot(C)$: Favors tracks with a constant velocity.
- $\theta_{Int} Int(C)$: Avoids collisions between targets.

❸ Proposed appearance model

Main idea:
Promotes configurations consistent with the **sparse representations** of the detections.



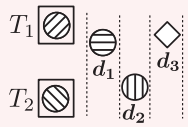
represented detection — track hypothesis — detection involved in the representation

$$App(C) = \sum_{\tau \in C} \sum_{d \in \tau} ||y_d - D_\tau \alpha^\tau_{y_d}||_2$$

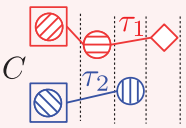
- y_d: feature vector of detection d
- D_τ: **dictionary** composed by elements of track τ
- α_{y_d}: **sparse representation** of detection d

$$\alpha_{y_d} = \arg\min_\alpha \frac{1}{2}||y_d - D\alpha||_2^2 + \lambda\Omega(\alpha)$$

- $D_\tau \alpha^\tau_{y_d}$: **residual reconstruction** over τ
- $||y_d - D_\tau \alpha^\tau_{y_d}||_2$: **residual reconstruction error**



$$D = [\,\,\,\,]$$
$$\alpha_{y_{d_3}} = [0 \;\; .3 \;\; .4 \;\; .2]^\top$$



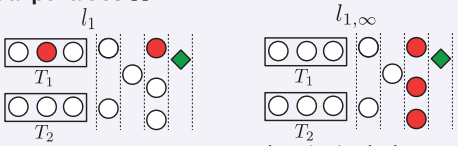
$$D_{\tau_1} = [\,\,\,]$$
$$\alpha^{\tau_1}_{y_{d_3}} = [.3 \;\; .4]^\top$$

$$App(C) = ||y_{d_1} - D_{\tau_1}\alpha^{\tau_1}_{y_{d_1}}||_2 + ||y_{d_2} - D_{\tau_2}\alpha^{\tau_2}_{y_{d_2}}||_2 + ||y_{d_3} - D_{\tau_1}\alpha^{\tau_1}_{y_{d_3}}||_2$$

❹ Structured sparsity

$$\alpha_{y_d} = \arg\min_\alpha \frac{1}{2}||y_d - D\alpha||_2^2 + \lambda\Omega(\alpha)$$

Usual penalties Ω:



l_1 norm: induces sparsity among all detections.

Issue: all detections of the target are not involved in the representation.

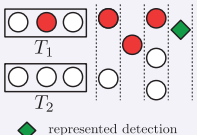
$l_{1,\infty}$ (or $l_{1,2}$) group norms: induce sparsity among groups of detections.

Issue: detections of other targets are involved in the representation.

Proposed penalty:

$$||\alpha||^w_{\infty,1} = \max_{i=1..\Delta t} w_i ||\alpha^{G_i}||_1$$

- Promotes representations where all groups are involved with a few elements inside each group.
- Frames as groups: favors to represent each detection with all the detections of the same target.

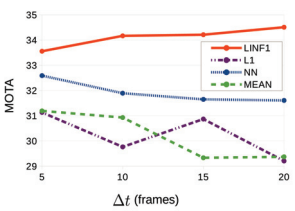


Sparsity structure induced by an $l_{\infty,1}$ norm.

Optimization:
- Exact optimization using proximal gradient descent methods (**FISTA**) with **active sets**.
- **Proximal operators** for a weighted $l_{\infty,1}$ norm efficiently evaluated with projections on the unit ball of the related dual norm.

❺ Results

Evaluation of some variants:
- Different sliding window sizes.
- Different appearance models.



Evaluation on the MOTChallenge 2015 training set.

→ Best results with $l_{\infty,1}$-based sparse representations.

Evaluation on the *MOTChallenge*:

MOTCh. 2015	MOTA	IDS	FM	FAF	MOTP
NOMT	33.7	442	823	1.3	71.9
MHT_DAM	32.4	435	826	1.6	71.8
MDP	30.3	680	1500	1.7	71.3
LP_SSVM	25.2	646	849	1.4	71.7
ELP	25.0	1396	1804	1.3	71.2
LINF1	24.5	298	744	1.0	71.3
JPDA_m	23.8	365	869	1.1	68.2
MotiCon	23.1	1018	1061	1.8	70.9
SegTrack	22.5	697	737	1.4	71.7
DCO_X	19.6	521	819	1.8	71.4
CEM	19.3	813	1023	2.5	70.7
RMOT	18.6	684	1282	2.2	69.6
SMOT	18.2	1148	2132	1.5	71.2
ALExTR.	17.0	1859	1872	1.6	71.2
TBD	15.9	1939	1963	2.6	70.9
GSCR	15.8	514	1010	1.3	69.4
TC_ODAL	15.1	637	1716	2.2	70.5
DP_NMS	14.5	4537	3090	2.3	70.8

MOT16	MOTA	IDS	FM	FAF	MOTP
LINF1	40.5	426	953	1.4	74.9
DP_NMS	31.9	969	941	0.2	76.4
SMOT	29.2	3072	4437	3.0	75.2

Results accessed on 14/03/2016, methods using public detections

***MOTChallenge 2015*:**
- Comparable results in **MOTA** with the majority of recent approaches.
- Best results in **IDS** and **FAF**, second best **FM**.
- 7.5 fps (4 cores CPU).

MOT16:
- Competitive with most of recent methods.

❻ Conclusion

- New energy formulation exploiting sparse representations for multi-frame data association.
- $l_{\infty,1}$**-based representations improve results** compared to l_1-based representations.
- **Robust tracking** in terms of **identity switches** and **track fragmentation**, comparing well with state-of-the-art approaches.
- **Future work:** extension to local features and joint sparse representation of all targets.

Contact: loic.fagot-bouquet@cea.fr