Deep Association Learning for Unsupervised Video Person Re-identification

Yanbei Chen[1] Xiatian Zhu[2] Shaogang Gong[1]

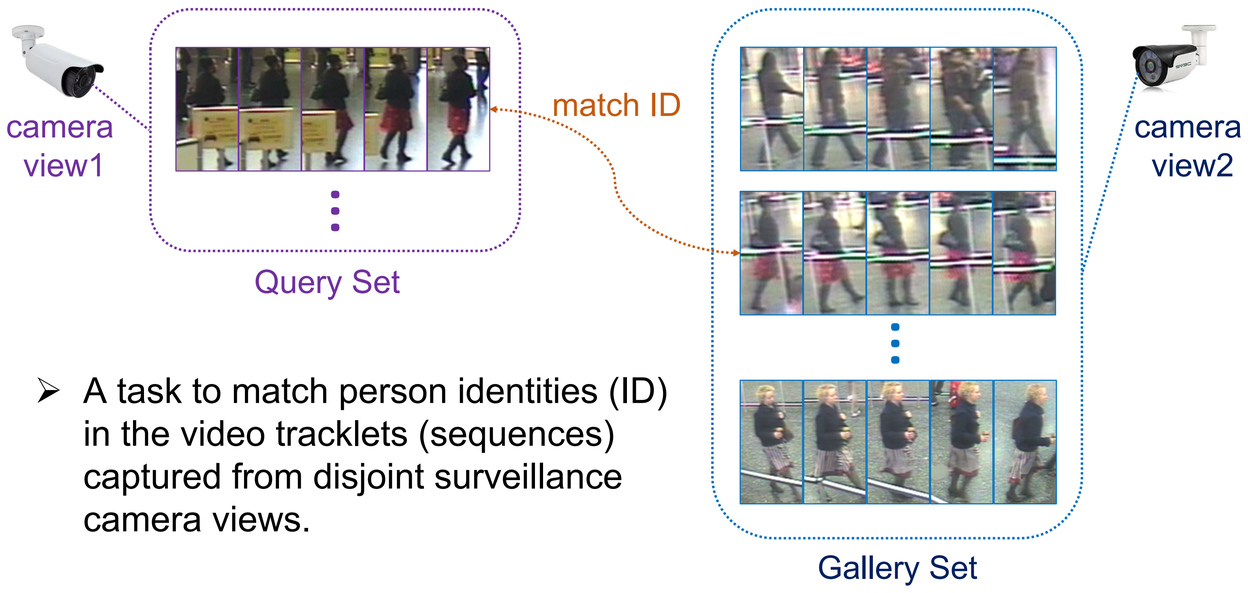
yanbei.chen@qmul.ac.uk eddy@visionsemantics.com s.gong@qmul.ac.uk

[1]Queen Mary University of London [2]Vision Semantics Ltd., London, UK

BMVC 2018 Newcastle upon Tyne

Introduction

Video Person Re-Identification (ReID)



- A task to match person identities (ID) in the video tracklets (sequences) captured from disjoint surveillance camera views.

Unsupervised Video Person ReID

❖ **Problem**

- How to formulate supervision signals without utilizing any pairwise ID matching information to guide model learning?

❖ **Main Contributions**

a) **A novel unsupervised learning method:** Deep Association Learning.
b) **End-to-end deep unsupervised learning framework** for video ReID: none manual labelled supervision is given for model training.
c) **State-of-the-art results** on various video ReID benchmark datasets.

Methodology Overview

Deep Association Learning

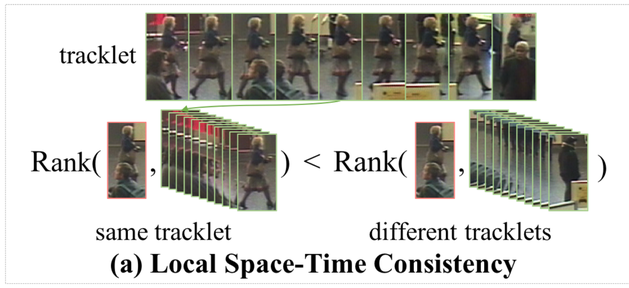
❖ **Overall idea**

- Learning by (1) intra-camera image-to-tracklet association and (2) cross-camera tracklet-to-tracklet association

Two types of association learning

1) Intra-Camera Association Learning

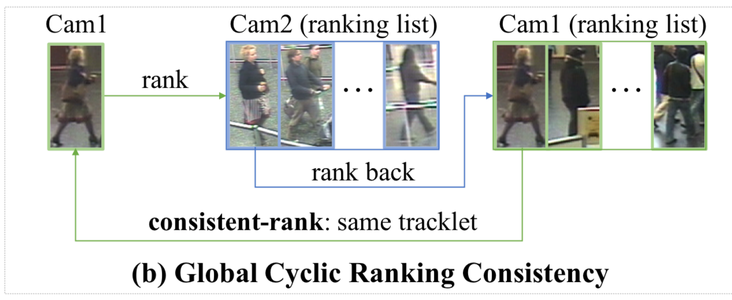
- image-to-tracklet association under the same camera view
- exploit the inherent label information for supervision



(a) Local Space-Time Consistency

2) Cross-Camera Association Learning

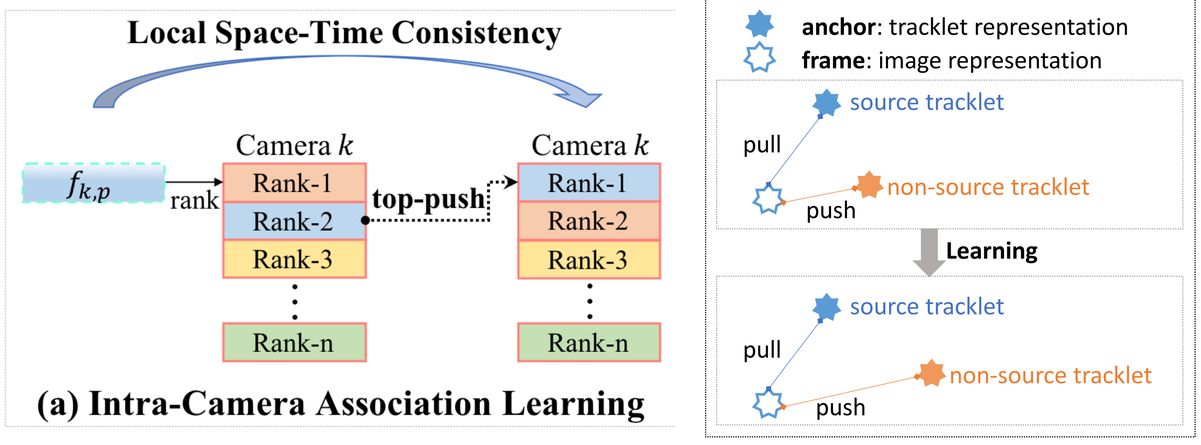
- tracklet-to-tracklet association across disjoint camera views
- mitigate the cross-camera domain gaps



(b) Global Cyclic Ranking Consistency

Intra-Camera Association Learning

Key Idea:
intra-camera image-to-tracklet association



(a) Intra-Camera Association Learning

Batch-wise three iterative steps

1) Learning Intra-Camera Anchors

Represent each tracklet as a learnable anchor

$$x_{k,i}^{t+1} \leftarrow x_{k,i}^{t} - \eta\left(\ell_2(x_{k,i}^{t}) - \ell_2(f_{k,p}^{t})\right), \text{ if } i = p$$

2) Tracklet Association Ranking

Rank all anchors in each camera view

$$\{D_{p,i} | D_{p,i} = \left\|\ell_2(f_{k,p}) - \ell_2(x_{k,i})\right\|_2, \ i \in N_k\}$$
$$\xrightarrow{\text{ranking}} D_{p,t} = \min_{i \in [1,N_k]} D_{p,i}$$

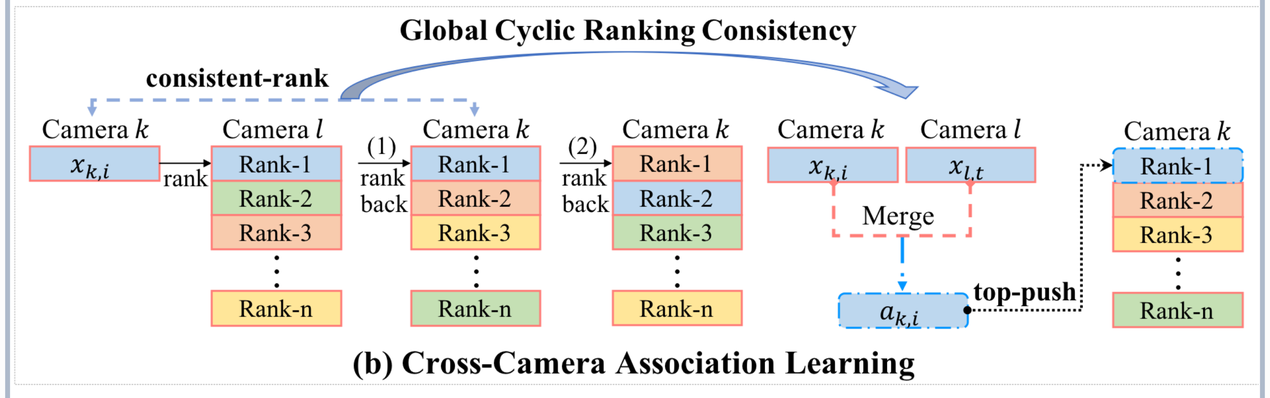
3) Intra-Camera Association Loss

Associate each frame with its own source tracklet

$$\mathcal{L}_I = \begin{cases} [D_{p,p} - D_{p,t} + m]_+, & \text{if } p \neq t \text{ (The rank-1 is not the source tracklet)} \\ [D_{p,p} - \overline{D_{j,t}} + m]_+, & \text{if } p = t \text{ (The rank-1 is the source tracklet)} \end{cases}$$

Cross-Camera Association Learning

Key Idea:
cross-camera tracklet-to-tracklet association



(b) Cross-Camera Association Learning

Batch-wise three iterative steps

1) Cyclic Ranking

Discover highly associated tracklets across disjoint camera views

$$x_{k,i} \xrightarrow{\text{ranking in cam } l} Dc_{p,t} = min_{i \in [1,N_l]} \xrightarrow{\text{ranking back in cam } k} Dc_{q,j} = min_{i \in [1,N_k]}$$

Global Cyclic Ranking Consistency: $j=i$

2) Learning Cross-Camera Anchors

Merge two highly associated tracklets as a cross-camera anchor

$$a_{k,i}^{t+1} \leftarrow \begin{cases} \frac{1}{2}\left(\ell_2(x_{k,i}^{t+1}) + \ell_2(x_{l,t}^{t})\right), & \text{if } j = i \quad \text{(Cyclic ranking consistent)} \\ x_{k,i}^{t+1}, & \text{others} \end{cases}$$

3) Cross-Camera Association Loss

Associate each frame with the best-matched cross-camera anchor

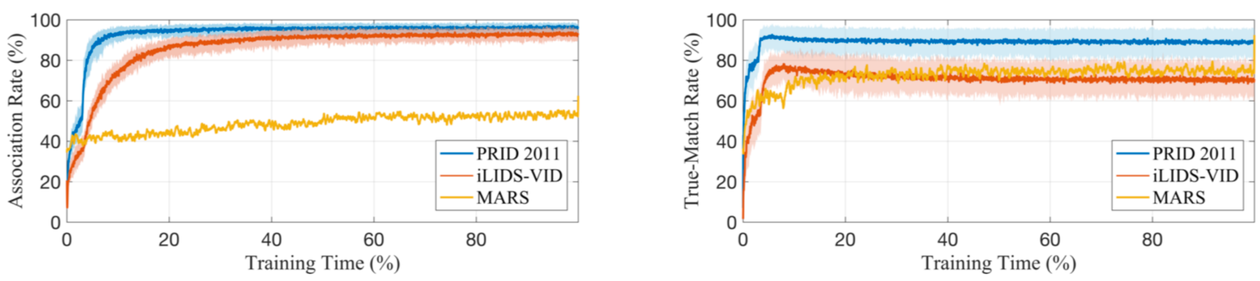
$$\mathcal{L}_C = \begin{cases} [Da_{p,p} - D_{p,t} + m]_+, & \text{if } p \neq t \text{ (The rank-1 is not the source tracklet)} \\ [Da_{p,p} - \overline{D_{j,t}} + m]_+, & \text{if } p = t \text{ (The rank-1 is the source tracklet)} \end{cases}$$

Experiments & Ablation Studies

➢ Comparison to state-of-the-art unsupervised video ReID methods

Datasets	PRID 2011				iLIDS-VID				MARS				
Rank@k	1	5	10	20	1	5	10	20	1	5	10	20	mAP
DVDL [18]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9	-	-	-	-	-
STFV3D [25]	42.1	71.9	84.4	91.6	37.0	64.3	77.0	86.9	-	-	-	-	-
MDTS-DTW [27]	41.7	67.1	79.4	90.1	31.5	62.1	72.8	82.4	-	-	-	-	-
UnKISS [19]	59.2	81.7	90.6	96.1	38.2	65.7	75.9	84.1	-	-	-	-	-
DGM+IDE [42]	56.4	81.3	88.0	96.4	36.2	62.8	73.6	82.7	36.8	54.0	61.6	68.5	21.3
Stepwise [26]	80.9	95.6	98.8	99.4	41.7	66.3	74.1	80.7	23.6	35.8	-	44.9	10.5
DAL (ResNet50)	85.3	97.0	98.8	99.6	56.9	80.6	87.3	91.9	46.8	63.9	71.6	77.5	21.4
DAL (MobileNet)	84.6	96.3	98.4	99.1	52.8	76.7	83.4	91.6	49.3	65.9	72.2	77.9	23.0

➢ Evolution of cross-camera tracklet association



(a) Evolution on association rate. (b) Evolution on true-match rate.