Video Based Person Reidentification As MDP

Arulkumar S (CS15D202)

December 13, 2018

Based on Paper: Jianfu Zhang, Naiyan Wang, and Liqing Zhang. Multi-shot pedestrian re-identification via sequentialdecision making. In *Computer Vision and Pattern Recognition (CVPR)*, 2018

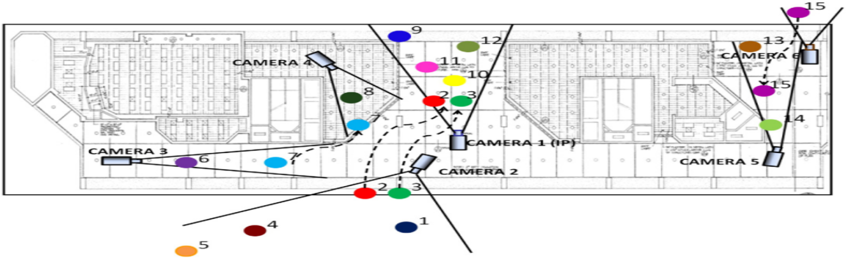
credits: `https://github.com/jiyanggao/Video-Person-ReID`

code: `https://github.com/InnovArul/personreid_sequential_rl`

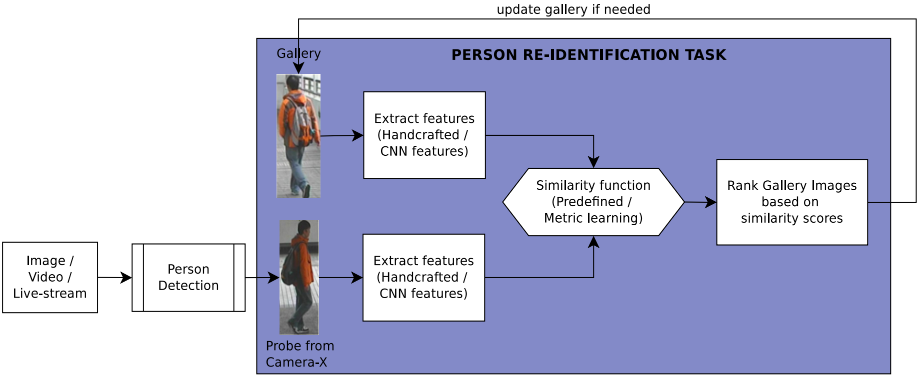
Person Re-Identification

Problem definition

- task of matching a person's image with images from database
- images captured at
  - same/different points in time (of same day)
  - same/different camera
  - various lighting conditions + unconstrained viewpoint/pose changes
- No information about camera position, intrinsic and extrinsic parameters



- **Probe:** the person's image to be searched in the database
- **Gallery:** one (or more) unique image(s) of persons observed so far. Usually, Gallery images will be available in a database.



- **Evaluation:** Ranking of matching scores (rank-1, rank-5, . . .)

MDP formulation

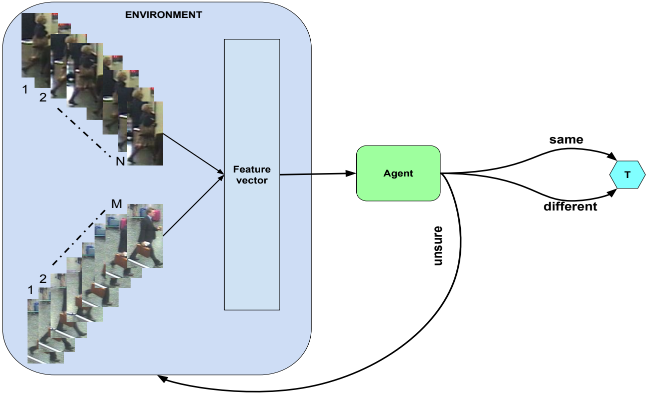


Figure: MDP formulation for the task of video based person re-identification. Here, *same, different, unsure* are actions.

States S_t

Let the features of video frames be $\{f_i\}_{i=1}^{N}, \{g_i\}_{i=1}^{N}$ and $f_i, g_i \in \mathbf{R}^d$

1. The absolute difference between the t^{th} frame descriptors (*d*-dimension), $o_t = |f_t - g_t|$
2. Historical features up-to time step t (*d*-dimension)

$$h_t = \begin{cases} o_t, & \text{if } t = 1 \\ \frac{\sum_{i=1}^{t-1} w_i \times o_i}{\sum_{i=1}^{t-1} w_i}, & \text{otherwise} \end{cases}$$

where,

$$w_i = 1 - \frac{e^{Q_u}}{e^{Q_s} + e^{Q_d} + e^{Q_u}}$$

3. $\texttt{mean}(\{||o_i||_{i=1}^{t}\}), \texttt{min}(\{||o_i||_{i=1}^{t}\}), \texttt{max}(\{||o_i||_{i=1}^{t}\})$ (3 dimension)

Total = $2d + 3$ dimensions

Actions A_t

- *same* - terminates immediately
- *different* - terminates immediately
- *unsure* - the feature corresponding to next image pair is given to the agent

Rewards

$$R_t = \begin{cases} +1, & \text{if } A_t \text{ matches ground truth (\textit{same / different}),} \\ -1, & \text{if } A_t \text{ doesn't match the ground truth (\textit{same / different}) or } (t = t_{max} \\ r_p, & t < t_{max} \text{ and } A_t \text{ is \textit{unsure}} \end{cases} \quad (1)$$

Base Network Architecture (Alexnet)

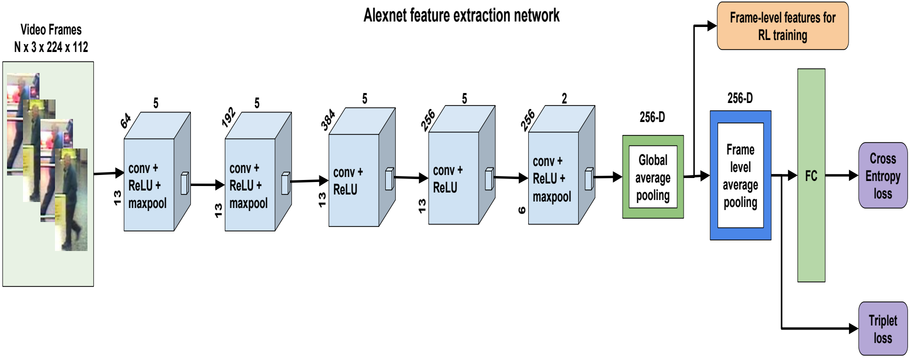


Figure: Alexnet base network used for pretraining and feature extraction for training RL-based DQN

Training losses

- Softmax Cross-Entropy Loss

$$L_{softmax} = -\frac{1}{P} \sum_{i} \sum t_j^i \log p_j^i \qquad (2)$$

- Triplet Loss

$$L_{triplet} = \sum_{i=1}^{P} \sum_{a=1}^{K} \max \left(m + \max_{p=1,\ldots,K} D(f_a^i, f_p^i) - \min_{n=1,\ldots,K} D(f_a^i, f_n^i), 0 \right)$$

- Total Loss

$$L = L_{softmax} + L_{triplet}$$

Downsides

- Features from all frames should be extracted and averaged

Downsides

- Features from all frames should be extracted and averaged
    - computationally inefficient!

Downsides

- Features from all frames should be extracted and averaged
    - computationally inefficient!
- The frames with severe occlusion might affect the **averaged** descriptor

DQN Architecture

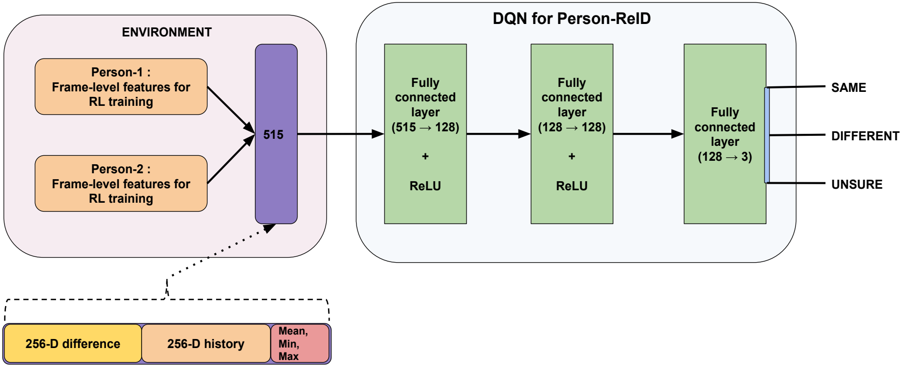


Figure: DQN architecture used for Person ReID Q-learning

DQN Training

- Replay Buffer (size = 10000, ϵ-Greedy exploration, ϵ reduces from 1 to 0.1)
- Target network (backup once in 10000 iterations)
- Q-learning

$$Q(s_t, a_t) = Q(s_t, a_t) + \eta(r_t + max_{a'_{t+1}} Q_m(s_{t+1}, a'_{t+1}) - Q(s_t, a_t))$$

Results in PRID2011 dataset

178 persons (train - 78, test - 78), 2 cameras, 100 frames per video on average

Method	mAP (%)	rank-1 (%)	rank-5 (%)	rank-10 (%)	rank-20 (%)	avg. #frames
RNN-CNN	-	70.00	90.00	95.00	97.00	100
Two stream	-	78.00	94.00	97.00	99.00	100
CNN +XQDA	-	77.9	93.5	-	99.3	100
baseline (alexnet, all frames)	86.2	80.9	93.3	96.6	100.0	100
baseline (alexnet, random 4 frames)	81.1	73.0	90.0	94.6	97.2	**4.00**
Q-learning ($r_p = 0.2$)	81.9	76.4	88.8	95.5	97.8	**3.949**
Q-learning ($r_p = 0.3$)	75.7	66.3	87.8	91.0	94.4	**5.05**

Table: Performance of the Q-learning based Person ReID method in PRID2011 dataset(**mAP** = Mean Average Precision (higher the value, better the method is), **rank-N** = ranked accuracy (higher the value, better the method is))

Quantitative results

Figure: Query (left), Gallery (right). Taken actions are *unsure, unsure, unsure, different*

Quantitative results

Figure: Query (left), Gallery (right). Taken actions are *unsure, different*

Quantitative results

Figure: Query (left), Gallery (right). Taken actions are *unsure, different*

Quantitative results

Figure: Query (left), Gallery (right). Taken actions are *unsure, same*

Observations

- RL method (8 frames) is slower than Baseline network with averaged features (8 frames)
  - Sequential nature of RL
  - Batch processing of episodes while testing? (not implemented in this project)

Observations

- RL method (8 frames) is slower than Baseline network with averaged features (8 frames)
  - Sequential nature of RL
  - Batch processing of episodes while testing? (not implemented in this project)
- Needs more hyper-parameter tuning (λ, ϵ, replay memory size etc.,) and training time was higher

Observations

- RL method (8 frames) is slower than Baseline network with averaged features (8 frames)
  - Sequential nature of RL
  - Batch processing of episodes while testing? (not implemented in this project)
- Needs more hyper-parameter tuning (λ, ϵ, replay memory size etc.,) and training time was higher
- Policy-gradient based approach (based on REINFORCE) did not converge/gives sub-optimal results (so far)

Questions?

Thank you!