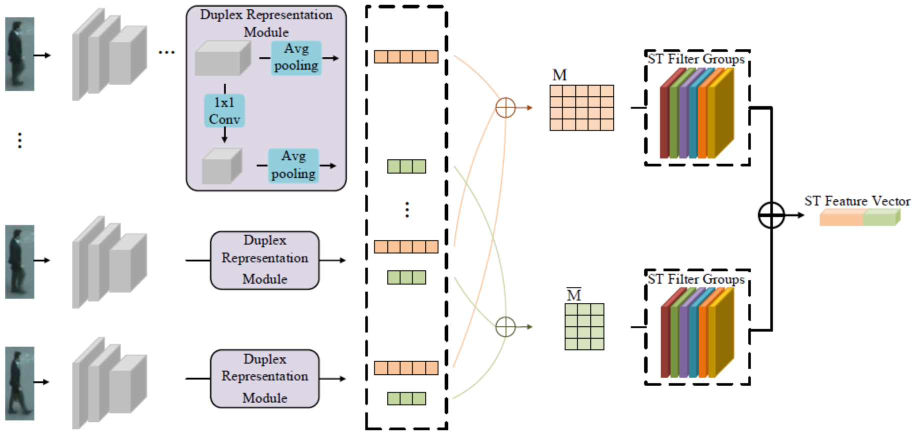
A Duplex Spatiotemporal Filtering Network for Video-based Person Re-identification

Chong Zheng, Ping Wei, Nanning Zheng

Overview

Person re-identification is to match pedestrians in different cameras and has gained increasing attention for its significant applications in areas such as video surveillance. However, it remains a challenging problem due to occlusion, background clutters, illumination variations, and camera viewpoint changes.

We propose a duplex spatiotemporal filtering network (DSFN) to jointly extract static and dynamic information from video sequences for person re-identification. The network structure are shown in the right figure.

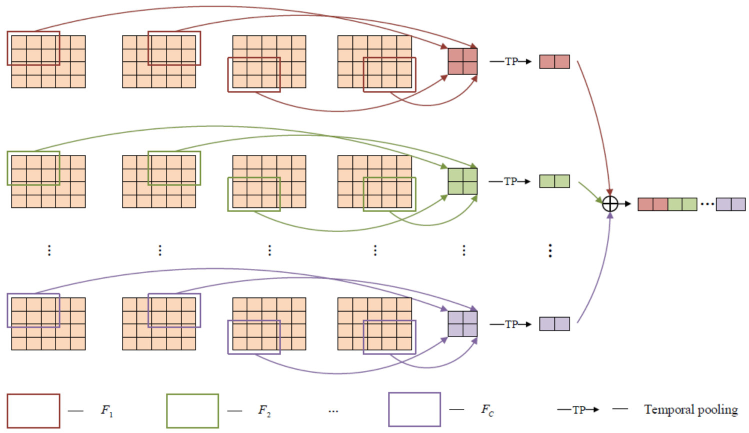


Spatial-Temporal Filter Groups

To extract spatial and temporal information from video sequences, we design novel filter groups. Compared with traditional filters, it has these advantages:

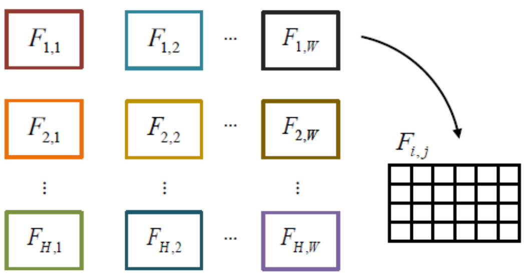
- Containing spatial and temporal dimension simultaneously
- Filter parameters are learned automatically other than fixed
- Special sparse-orthogonal constrains are designed to enhance the performance of filter groups.

The filtering procedure is shown in the right figure.



Sparse-Orthogonal Loss Function

We turn sparse-orthogonal constrains into loss function to actually realize the constrains. Here is an example picture of filter groups:



Orthogonal limitation between spatial filter groups:

$$L_O = \frac{1}{Q}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{j'=1}^{W} P(F_{i,j}, F_{i,j'})$$

Sparse limitation between temporal filter groups:

$$L_D = \frac{1}{Q}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{j'=1}^{W} D(F_{i,j}, F_{i,j'})$$

Sparse-Orthogonal Loss Function :

$$L_F = L_O - \alpha L_D$$

Loss Function of the whole training process:

$$L = L_C + L_R + \beta L_F$$

Results and Visualization

1.PRID2011

Methods	rank-1	rank-5	rank-20
TDL[19]	56.7	80.0	93.6
MARS[23]	53.0	81.4	95.1
SeeForest[9]	79.4	94.4	99.3
QAN[30]	90.3	98.2	97.4
RQEN[31]	91.8	98.4	99.8
DRSA[10]	93.2	-	-
COSAM[7]	79.6	95.3	-
M3D[32]	94.4	**100.0**	-
snippet[12]	93.0	99.3	**100.0**
DSFN	**96.0**	98.7	99.7

2.iLIDS-VID

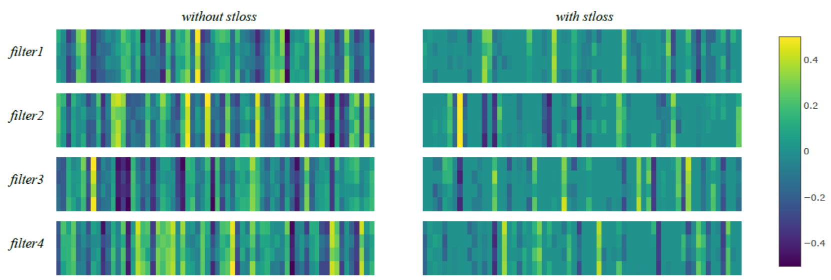
Methods	rank-1	rank-5	rank-20
TDL[19]	56.3	87.6	98.3
MARS[23]	53.0	81.4	95.1
SeeForest[9]	55.2	86.5	97.0
QAN[30]	68.0	86.8	97.4
RQEN[31]	77.1	93.2	99.4
DRSA[10]	80.2	-	-
COSAM[7]	79.6	95.3	-
M3D[32]	74.0	94.3	-
snippet[12]	85.4	96.7	99.5
DSFN	**87.7**	**97.0**	**99.6**

3.MARS

Methods	mAP	rank-1	rank-5	rank-20
MARS[23]	49.3	68.3	82.6	89.4
SeeForest[9]	-	70.6	90.0	97.6
QAN[30]	51.7	73.7	84.9	91.6
RQEN[31]	71.1	77.8	88.8	94.3
DRSA[10]	65.8	82.3	-	-
TriNet[22]	67.7	79.8	91.4	-
K-reciprocal[33]	68.5	73.9	-	-
M3D[32]	74.1	84.4	-	97.8
snippet[12]	76.1	86.3	94.7	**98.2**
COSAM[7]	**79.9**	84.9	95.5	98.0
DSFN	79.5	**86.6**	**95.8**	97.8
DSFN+Re-ranking	**87.4**	**87.8**	94.9	97.8

Ablation experiment result

Dataset	iLIDS-VID		MARS	
Methods	rank-1	rank-5	mAP	rank-1
Baseline	84.5	96.5	78.3	84.8
STF	85.1	96.3	78.9	85.7
STF+DR	85.9	96.6	79.4	86.3
STF+DR+FC	**87.7**	**97.0**	**79.5**	**86.6**



The visualization of filter groups. After the spatio-temporal loss function is added to the training, the value of more filters is limited to 0, which realizes the sparseness limitation of the loss function, so that different filter groups can focus on extracting different features