Person Re-Identification via Recurrent Feature Aggregation

Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, Xiaokang Yang

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

UNIVERSITY OF MICHIGAN

Introduction

- We address the person re-identification problem by exploiting a globally feature representation from a sequence of tracked human regions/patches;
- We show that a progressive fusion framework based on LSTM aggregates the frame-wise human region representation and yields a sequence level feature representation;
- Experimental results on two person re-identification benchmarks demonstrate that the proposed method performs favorably against state-of-the-art person re-identification methods.

Recurrent Feature Aggregation Framework

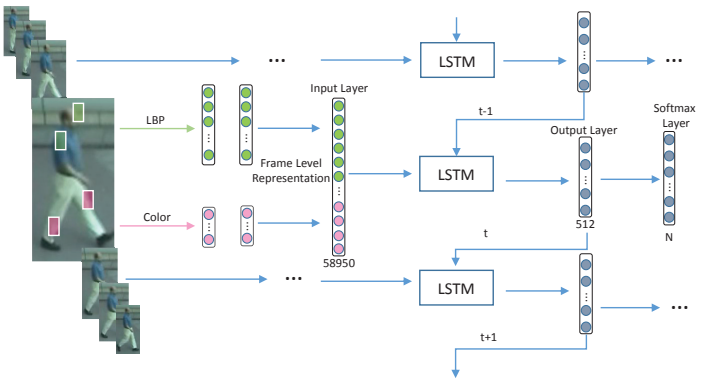

Fig. 1. Detailed structure of the framework based on LSTM.

- LBP and Color features are first extracted from rectangular image patches, and then concatenated as frame level representation.
- This representation and the previous LSTM outputs are input to the current LSTM node.
- At each time stamp t, given the input $\mathbf{x}_t$ and the previous hidden state $\mathbf{h}_{t-1}$, we update the LSTM network as follows:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i\mathbf{x}_t + \mathbf{U}_i\mathbf{h}_{t-1} + \mathbf{V}_i\mathbf{c}_{t-1} + \mathbf{b}_i) \quad (1)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f\mathbf{x}_t + \mathbf{U}_f\mathbf{h}_{t-1} + \mathbf{V}_f\mathbf{c}_{t-1} + \mathbf{b}_f) \quad (2)$$

$$\mathbf{c}_t = \mathbf{f}_t \cdot \mathbf{c}_{t-1} + \mathbf{i}_t \cdot \tanh(\mathbf{W}_c\mathbf{x}_t + \mathbf{U}_c\mathbf{h}_{t-1} + \mathbf{b}_c) \quad (3)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o\mathbf{x}_t + \mathbf{U}_o\mathbf{h}_{t-1} + \mathbf{V}_o\mathbf{c}_t + \mathbf{b}_o) \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t \cdot \tanh(\mathbf{c}_t) \quad (5)$$

- The output of the LSTM hidden state $\mathbf{h}_t$ is further connected to a softmax layer. The output of the N-way softmax is the prediction of the probability distribution over N different identities.

$$y_i = \frac{\exp(y_i')}{\sum_{k=1}^{N}\exp(y_k')}, \quad (6)$$

- The network is learned by minimizing $-\log y_k$, where k is the index of the true label for a given input. Stochastic gradient descent is used with gradients calculated by back-propagation.

Experimental Results

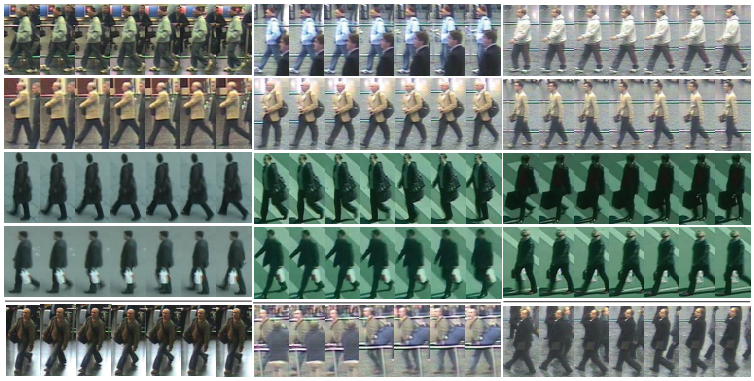

Fig. 2. Matching examples. (a) probe sequence, (b) Correct match, (c) Rank-1 matching sequence. The first four rows correspond to failure examples. The bottom row illustrates an example that fails when using the features from the first LSTM node, but succeeds when using the features accumulated to the 10th LSTM node.

Table 1. Performance of different methods based on Color&LBP feature

Dataset	iLIDS-VID				PRID 2011			
Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
Color&LBP [1]+RSVM	23.2	44.2	54.1	68.8	34.3	56.0	65.5	77.3
Color&LBP+DTW [2]	9.3	21.7	29.5	43	14.6	33	42.6	47.8
Color&LBP+DVR [3]	34.5	56.7	67.5	77.5	37.6	63.9	75.3	89.4
Color&LBP+**RFA-Net**+Cosine	44.5	71.9	82.0	**90.1**	54.9	84.2	**93.7**	**98.4**
Color&LBP+**RFA-Net**+RSVM	**49.3**	**76.8**	**85.3**	90.0	**58.2**	**85.8**	93.4	97.9

Table 2. Performance of our method in existence of noises

Dataset	iLIDS-VID				PRID 2011			
Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
Noise Level: 0%	49.3	76.8	85.3	90.0	58.2	85.8	93.4	97.9
Noise Level: 10%	43.4	70.6	81.5	88.9	52.3	83.2	91.4	97.5
Noise Level: 30%	40.0	67.4	77.5	87.0	51.4	81.1	90.5	96.9
Noise Level: 50%	29.8	60.5	71.9	81.5	44.7	75.2	85.6	95.5

Table 3. Performance of our method compared against state-of-the-art methods.

Dataset	iLIDS-VID				PRID 2011			
Rank R	R=1	R=5	R=10	R=20	R=1	R=5	R=10	R=20
3D Hog&Color+DVR [3]	39.5	61.1	71.7	81.0	40.0	71.7	84.5	92.2
DVDL	25.9	48.2	57.3	68.9	40.6	69.7	77.8	85.6
STFV3D [4]	37.0	64.3	77.0	86.9	21.6	46.4	58.3	73.8
STFV3D+KISSME	44.3	71.7	83.7	**91.7**	**64.1**	**87.3**	89.9	92.0
Color&LBP+**Our**+Cosine	44.5	71.9	82.0	90.1	54.9	84.2	**93.7**	**98.4**
Color&LBP+**Our**+RSVM	**49.3**	**76.8**	**85.3**	90.0	58.2	85.8	93.4	97.9

References

[1] Hirzer, M., Roth, P.M., Kostinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: ECCV. (2012)

[2] Simonnet, D., Lewandowski, M., Velastin, S.A., Orwell, J., Turkbeyler, E.: Re-identification of pedestrians in crowds using dynamic time warping. In: ECCV. (2012)

[3] Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: ECCV. (2014)

[4] Liu, K., Ma, B., Zhang, W., Huang, R.: A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In: ICCV. (2015)

More information at https://sites.google.com/site/yanyichao91sjtu