The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Temporal-Enhanced Convolutional Network for Person Re-identification

Yang Wu, Jie Qiu, Jun Takamatsu, Tsukasa Ogasawara
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara, 630-0192, Japan
yangwu@rsc.naist.jp, {qiu.jie.qf3, j-taka, ogasawar}@is.naist.jp

Abstract

We propose a new neural network called Temporal-enhanced Convolutional Network (T-CN) for video-based person re-identification. For each video sequence of a person, a spatial convolutional subnet is first applied to each frame for representing appearance information, and then a temporal convolutional subnet links small ranges of continuous frames to extract local motion information. Such spatial and temporal convolutions together construct our T-CN based representation. Finally, a recurrent network is utilized to further explore global dynamics, followed by temporal pooling to generate an overall feature vector for the whole sequence. In the training stage, a Siamese network architecture is adopted to jointly optimize all the components with losses covering both identification and verification. In the testing stage, our network generates an overall discriminative feature representation for each input video sequence (whose length may vary a lot) in a feed-forward way, and even a simple Euclidean distance based matching can generate good re-identification results. Experiments on the most widely used benchmark datasets demonstrate the superiority of our proposal, in comparison with the state-of-the-art.

Introduction

Person re-identification (Re-ID) is about associating tracks of people by their identities from potentially any number of cameras with or without overlapping views. Re-ID plays a very important role in intelligent video surveillance, as it links lower-level analytic results, such as human detection and tracking, with higher-level demands, such as tracing specific persons (e.g. suspects) and understanding human activities in a relatively large time and space scale (e.g. one day's activity at different places). However, it is not easy to solve this problem, because Re-ID requires reliable association of the same person given possibly large appearance changes due to body motion, interactions with objects and/or other people, dynamic and cluttered backgrounds, viewpoint and environmental changes between cameras, etc. Meanwhile, Re-ID needs to ensure good differentiation among people with similar appearances (body shape, clothes, etc.), which is hard when the resolution is low.

Existing studies on person Re-ID fall into two categories based on the type of input data: **image-based person Re-ID** when the inputs are individual still images, and **video-based person Re-ID** when inputting video sequences (Zheng, Yang, and Hauptmann 2016). The later is a relatively less popular topic, because of some additional challenges such as arbitrary sequence lengths for each person, unknown dynamic occlusions, extra noises brought by tracking inaccuracy, and a lack of public benchmarks in the early dates due to hardness for ground-truth labeling. However, the most typical application of person Re-ID is video surveillance, for which captured data is usually in the form of videos, so it is more natural to directly work on videos. Moreover, videos contain not only much richer spatial information (about appearance) than few still images but also important temporal information (e.g. gait) that does not exist in individual images but can be very helpful for re-identification, so using videos as input can generally result in better performance.

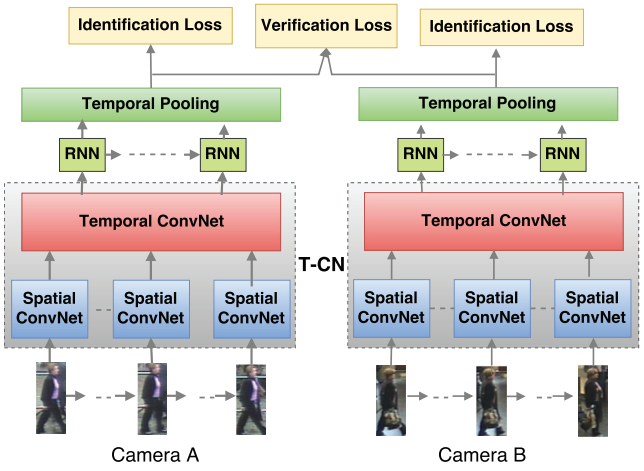

Figure 1: The overall architecture of our proposed model.

This paper contributes to video-based person Re-ID by proposing a new neural network called Temporal-enhanced Convolutional Network (T-CN), as shown in Figure 1. The main contributions and novelties are two-fold (*more details are given in the following related work section*).

1. **Introducing a novel network structure based on temporal convolutions (thus it is called *Temporal ConvNet*) for low-level and/or mid-level motion representation.**

Low-level and mid-level motion representation is critical to temporal information modeling but unfortunately there is a lack of in-depth and proper solutions due to its hardness. T-CN goes beyond the existing idea of exploring optical flow and temporal pooling (McLaughlin, Martinez del Rincon, and Miller 2016) by introducing a simple and light-weight temporal convolution block that can work much better.

2. **Proposing an end-to-end deep neural network architecture for effective and robust spatial and temporal representation learning.** The proposal integrates Temporal ConvNet with image-based convolutional network (denoted by *Spatial ConvNet* here), recurrent neural networks (RNN) and temporal pooling, which are all justified for video-based person Re-ID in the latest references. A Siamese network structure is adopted for optimizing both identification loss and verification loss. The whole network can take raw video sequences as input and be trained end-to-end using backpropagation through time. In testing, the network works for video sequences with various lengths and directly output the final feature vectors which are ready for Euclidean distance based ranking (no need further mapping or learning).

In the view of functionality, Spatial ConvNet models appearance (spatial information), Temporal ConvNet works on lower-level (low- and mid-level) motion (local temporal information), and finally RNN and temporal pooling extracts higher-level dynamics (global temporal information). We demonstrate the effectiveness and generalization ability of our proposal on two commonly used benchmark datasets, showing that it outperforms the state-of-the-art.

Related work

Image-based person re-identification

Image-based person Re-ID has a very rich literature (Zheng, Yang, and Hauptmann 2016). Traditional solutions generally follow the hand-crafted appearance feature description plus distance metric framework. Representative and relatively more influential features include the color and texture histograms over partitioned horizontal stripes (Gray and Tao 2008), the local maximal occurrence (LOMO) descriptor (Liao et al. 2015) and the hierarchical Gaussian descriptor (Matsukawa et al. 2016). Besides appearance features, semantic attribute based representations have also been investigated recently (Layne et al. 2012; Shi, Hospedales, and Xiang 2015; Li et al. 2016), which is intuitive and consistent with human knowledge but hard to extract due to the large semantic gap between image data and attributes. Since hand-crafted features are usually not discriminative enough for direct Euclidean distance based ranking, much work has been done on distance metric learning, for which the KISSME (Keep It Simple and Straightforward MEtric) model (Köstinger et al. 2012) and the XQDA (Cross-view Quadratic Discriminant Analysis) model (Liao et al. 2015) are most influential representatives. Dimension reduction models such as pairwise constrained component analysis (PCCA) (Mignon and Jurie 2012) and kernel LFDA (Pedagadi et al. 2013) have also been investigated for further enhancing the performance. Recently, deep learning has been successfully introduced to image-based Re-ID and gets popularized very quickly, with performances going beyond traditional models (Zheng, Yang, and Hauptmann 2016).

Video-based person re-identification

Research on video-based person Re-ID is much more diverse than that on image-based setting. Though spatial information is still very important, temporal information can be also very useful. All recent representative works adopt spatial appearance information, so here we categorize them by the way they explore temporal information. Roughly, we can put them into four groups.

The first group do not use any temporal information. Instead, they explore various ways to make use of individual video frames. A straightforward way is to explore discriminative learning (e.g. subspace learning) on them (Li et al. 2015). Similarly, another way is to train a multiple class classifier (e.g. using a deep CNN model) and then do simple temporal pooling over all the frames in a video sequence to get an overall vector-based representation, which may be fed into a metric learning model, as done in (Zheng et al. 2016). Much work has also been done on directly treating each sequence as a set of images and explore set-based recognition models, such as those based on set-to-set distances (Wu et al. 2012), and sparse representation with dictionary learning (Wu et al. 2015; Karanam, Li, and Radke 2015). Though encouraging results have been got, they miss all the benefits that temporal information may bring.

The second group extracts hand-crafted spatial-temporal features (or signatures) without explicit alignment of walking circle. Different approaches on using these spatial-temporal features have been tested. The simplest is to directly use the features for unsupervised matching (Hamdoun et al. 2008; Farenzena et al. 2010), and there is also further enhancement with metric learning (You et al. 2016). More sophisticated ones even have automatic discriminative video fragment selection (Wang et al. 2016) which is a kind of rough temporal alignment or rough body pose classification which can be regarded as rough spatial alignment. However, these alignment steps need extra efforts to make sure the imperfect alignment brings more benefits than hurts.

The third group relies on precise temporal alignment for feature representation. The alignment is usually done with an independent and explicit walking circle extraction model, e.g. Flow Energy Profile (FEP) (Liu et al. 2015). The temporal aligned features can be spatio-temporal body-action primitives encoded by 3D Fisher vectors (Liu et al. 2015) or just local average pooling of frame-wise appearance features like LOMO features (Gao et al. 2016). Then re-identification is done by applying existing metric learning models or a customized new model (Zhu et al. 2016) to these features. This group make the best use of temporal information, but it highly relies on a quality alignment which is hard to ensure.

The last group just includes an impressive recent work on exploring a new deep learning architecture called recurrent convolutional network (denoted by *RCN* in this paper) for a joint learning of spatial and temporal representations by link convolutional neural network (CNN), recurrent neural net-

work (RNN) and temporal pooling together (McLaughlin, Martinez del Rincon, and Miller 2016). It is claimed that the network covers three levels (low, middle, and high) of temporal information via optical flow, RNN and temporal pooling, respectively. However, we believe that both RNN and temporal pooling are better at the high-level abstraction and the optical flow extraction for low-level temporal information is an independent step out of the control by the network. Our proposed network T-CN inherits the merits of RCN and adds the low-level and mid-level temporal representation learning by introducing the temporal concolution network. Our experiments to be shown later verify that T-CN can model temporal information better than RCN, while at the same time it can replace optical flow's ability with almost no additional cost.

Temporal-enhanced convolutional network

Convolution networks are a kind of well-known neural networks for representation learning, especially when the input is 2D image data. Such a network generally contains multiple convolution layers, in combination with pooling and nonlinear mapping layers, for multi-scale representation and abstraction. We extend the idea of convolution for extracting both spatial information and temporal information, in the way of 2D spatial convolution over images and 1D temporal convolution over time, respectively.

As shown in the overall architecture of our model in Figure 1, the raw video frames of each sequence (capturing one specific person under a certain camera) are first fed into a convolutional network called Temporal-enhanced Convolutional Network (T-CN) for feature extraction. Such a network contains two concatenating subnets: spatial convolution network and temporal convolution network, where the former focuses on extracting multi-scale spatial information for appearance modeling and the later further explores low level and middle level temporal information for motion modeling by linking local neighboring frames.

The output of T-CN is a sequence of feature vectors, each of which abstracts the appearance and up to mid-level motion information of several consecutive video frames. These feature vectors are then further fed into a Recurrent Neural Network (RNN) and a temporal pooling layer for overall appearance and dynamics extraction. Therefore, T-CN is of great importance for all-level temporal information modeling and integration.

Spatial Convolution Network

The spatial convolution network (*Spatial ConvNet*) is a neural network with multiple convolution layers, as shown in Figure 2 with detailed parameters stated. The first layer takes an image as its input. To be generic, each channel of the image can be regraded as a feature map. The first layer applies convolution to all these feature maps (image channels) with a set of convolution kernels, producing new feature maps. Subsequential convolution layers operate similarly, just having their inputs coming from the feature maps generated by their previous layers. After each convolution layer, there is max pooling layer and a nonlinear hyperboic-tangent (Tanh) activation (not shown in the Figure). Note that in the T-CN architecture, all the frames are fed into the same Spatial ConvNet, or in other words, the Spatial ConvNet share the same parameters across all the input frames.

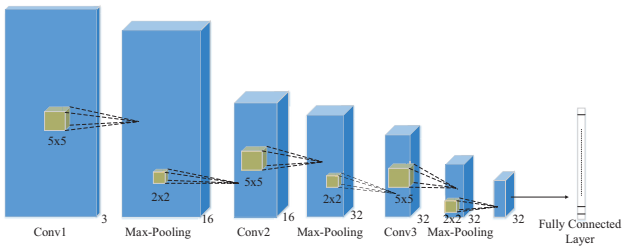

Figure 2: Spatial Convolution Network.

For a brief formulation, let $\mathbf{F} = \{F_1, \ldots, F_n\}$ denote an input video sequence with n frames, where $F_i, i \in \{i, \ldots, n\}$ indicates the frame i. For simplicity, we use C_s to stand for the function of the whole Spatial ConvNet, then we can get the output vector $\mathbf{v}_i^s \in \mathbb{R}^d$ in d-dimensional space for F_i by

$$\mathbf{v}_i^s = C_s(F_i). \tag{1}$$

Temporal Convolution Network

Our temporal convolution network targets at extracting low-level and mid-level temporal information. To do that, we build such a network with two layers: temporal convolution layer and kernel response integration layer, as shown in Figure 3. The former one does a 1D convolution along the time axis with a set of totally K kernels, which generates K different responses. Then we have another layer integrate all the responses into one single vector for each output element, using a full-length 1D convolution with a kernel sized $1 \times K$.

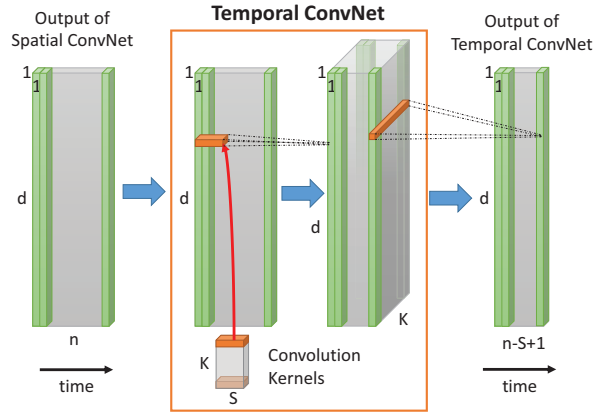

Figure 3: Temporal Convolution Network.

Suppose the output sequence of Spatial ConvNet for the whole video sequence can be denoted by a matrix $\mathbf{V}^s \in \mathbb{R}^{d \times n}$, and let C_t to denote the function of Temporal ConvNet, and we will have

$$\mathbf{V}^t = C_t(\mathbf{V}^s). \tag{2}$$

where $\mathbf{V}^t \in \mathbb{R}^{d\times(n-S+1)}$ is the output sequence, with each column indicating an element.

In greater details, the two different layers within Temporal ConvNet can be modeled as follows.

$$\mathbf{x}_j^k = tanh(\sum_{u=1}^{S} \mathbf{w}_{TC}^k(u)\,\mathbf{v}_{j+u-1}^s), \forall j \in \{1,\ldots,n-S+1\}.$$

$$\mathbf{v}_j^t = tanh(\sum_{k=1}^{K} \mathbf{w}_{FI}(k)\,\mathbf{x}_j^k), \forall j \in \{1,\ldots,n-S+1\}.$$

where $\mathbf{w}_{TC}^k, k \in \{1,\ldots,K\}$ and $\mathbf{w}_{FI}$ denote the convolution kernels for temporal convolution and kernel response integration, respectively, and their u-th and k-th elements are indicated by (u) and (k). Note that each convolution is followed by a $tanh$ non-linear mapping.

High-level motion modeling

RNN is used for extracting high-level motion information. Since the output of RNN is still a sequence of elements in the time order, with each element aggregating information up to the current time point (closer ones are aggregated more), an extra temporal pooling step integrates all the information together into a sequence level representation.

Recurrent neural network

Recurrent neural networks, known as RNNs, are a family of neural networks for processing sequential data. Most RNNs could deal with long sequences with variable lengths. For each time step, the structure of RNNs share the same parameters, which enable extending and applying the model to samples with different lengths. We use the same simple RNN structure as introduced in RCN, as it is proved to be suitable for our task by experiment. Since RNN is known to be good at sequence-level dynamic modeling and its effectiveness usually require that the input contains clear patterns over time, we think it is better to have some other network extracting up to mid-level representation (not only for the spatial information but also for the temporal information) before feeding into RNN. Therefore, our proposed T-CN contains temporal convolution network for that purpose.

For the input sequence $\mathbf{v}_j^t, j \in \{1,\ldots,n-S+1\}$, RNN does the following:

$$\mathbf{v}_j^r = \mathbf{v}_j^t + \mathbf{W}_{RNN}^T \mathbf{r}_{j-1}, \tag{3}$$

$$\mathbf{r}_j = tanh(\mathbf{v}_j^r), \tag{4}$$

where $\mathbf{W}_{RNN} \in \mathbb{R}^{d\times d}$ is the projection matrix of RNN, $tanh$ is the non-linear hyperbolic-tangent activation function, $\mathbf{r}_j$ is the RNN state at step j, and $\mathbf{v}_j^r$ is the output of RNN for element j. $mathbf{r}_0$ is initialised to the zero-vertor for computation.

Temporal Pooling

Temporal pooling over RNN outputs can aggregate information over the whole sequence and avoid RNN's bias towards later time-steps. It has been proved to be a simple yet effective way to represent gait information, for which GEI (Han and Bhanu 2006) is a good example. Temporal pooling is alignment free, as itself is totally independent from the order of the sequence. There exist two widely used types of temporal pooling: average-pooling and max-pooling, among which average-pooling is empirically proved to be more suitable for aggregation the RNN outputs (McLaughlin, Martinez del Rincon, and Miller 2016). Therefore, we choose average pooling for our model.

Given the output $\mathbf{v}_j^r, j \in \{1,\ldots,n-S+1\}$ of RNN, the average-pooling generates the final feature vector $\mathbf{v}^o \in \mathbb{R}^d$ by

$$\mathbf{v}^o = \frac{1}{n-S+1}\sum_{j=1}^{n-S+1} \mathbf{v}_j^r. \tag{5}$$

Model training by joint identification and verification

Person re-identification is commonly treated as a verification task, as generally Re-ID does not need to tell who the probe person is, but to look for the corresponding person in the gallery. For deep learning models, this is usually achieved by optimizing a verification loss with a Siamese network architecture (Hadsell, Chopra, and LeCun 2006) as shown in Figure 1. Siamese network is a symmetric structure, containing two parts that share all the parameters. Generally, one sequences from one camera and the other from different camera should be fed into two corresponding parts of Siamese network. Depending on the labels of these sequences, the Siamese structure would perform different activities based on the verification loss. If the input is about the same person, the verification loss should be small and the Siamese network should try to make the learned features closer, while for the opposite case (given sequences from different persons), the verification loss should be high and the Siamese network would separate them towards a prefixed margin. Besides verification loss, identification loss is also applicable for deep models with data augmentation, and it has already been proved to be effective. So here we jointly optimize both losses as shown in Figure 1.

Suppose there are c different persons and an input sequence whose final feature vector is $\mathbf{v}^o \in \mathbb{R}^d$ with identity $y \in \{1,\ldots,c\}$, the **identification loss** $L_I(\mathbf{v}^o)$ for this sequence based on softmax function can be defined as follows:

$$L_I(\mathbf{v}^o) = P(y=l|\mathbf{v}^o) = \frac{exp(\mathbf{W}_{SM,l}^T\mathbf{v}^o)}{\sum_{q=1}^{c} exp(\mathbf{W}_{SM,q}^T\mathbf{v}^o)} \tag{6}$$

where $\mathbf{W}_{SM,l}$ and $\mathbf{W}_{SM,q}$ are the l^{th} and q^{th} column of the softmax weight matrix $\mathbf{W}_{SM}$, respectively.

Given a pair of sequences with feature vectors $\mathbf{v}_a^o \in \mathbb{R}^d$ and $\mathbf{v}_b^o \in \mathbb{R}^d$, and with identities y_a and y_b, respectively, then the **verification loss** $L_V(\mathbf{v}_a^o, \mathbf{v}_b^o)$ can be defined as:

$$L_V(\mathbf{v}_a^o, \mathbf{v}_b^o) = \begin{cases} \frac{1}{2}\left\|\mathbf{v}_a^o - \mathbf{v}_b^o\right\|_2^2, & y_a = y_b \\ \frac{1}{2}[max\left(m - \left\|\mathbf{v}_a^o - \mathbf{v}_b^o\right\|, 0\right)]^2, & y_a \neq y_b \end{cases} \tag{7}$$

The all overall loss is a sum of these two types of losses:

$$L(\mathbf{v}_a^o, \mathbf{v}_b^o) = L_I(\mathbf{v}_a^o) + L_I(\mathbf{v}_b^o) + L_V(\mathbf{v}_a^o, \mathbf{v}_b^o). \tag{8}$$

Experiments

Datasets and implementation details

We evaluate our proposed model (T-CN) on two commonly used benchmark datasets: PRID 2011 (Hirzer et al. 2011) and iLIDS-VID (Wang et al. 2014), for a fair comparison with the state-of-the-art and a detailed empirical study of our proposal.

PRID 2011 dataset. The PRID 2011 dataset contains data from two cameras with non-overlapping views, where camera A captured 385 persons and camera B captured 749 persons. Only the first 200 persons from each of them appear in both camera views. The literature so far have commonly focused on matching the 200 common persons with one single video sequence for each person from each camera view (Wang et al. 2014). Totally there are 400 sequences and each sequence has a variable length of 5 to 675 frames. Here we use the same setting for a fair comparison, though the dataset builders suggested three different evaluation settings with gallery set(s) covering unseen people. Many existing models have only been evaluated on a selected subset of 178 people (89% of the data) as they require longer sequences (e.g. longer than 21 frames) for being effective (details are shown in Table 1). Our proposed model doesn't have such a limitation, so we ran our experiments on all the 200 people. It is worth mentioning that the data were collected in an uncrowded outdoor environment with relatively simple and clean background and serious occlusions have been filtered out. The main challenges are viewpoint and illumination changes between cameras. Therefore, this dataset is good for idea verification, and usually relatively good performance can be achieved.

iLIDS-VID dataset. The iLIDS-VID dataset also has only two cameras with non-overlapping views. It has slightly larger size: 600 video sequences for 300 people. Again, each person has only one sequence from each camera. The sequence length varies from 23 to 192, with an average number of 73. Unlike PRID 2011 dataset, the iLIDS-VID data comes from a busy airport arrival hall, so it contains crowed background clutter, heavy occlusions, and significant viewpoint/illumination variations, making the dataset very challenging. This dataset is good for effectiveness justification.

Implementation details of T-CN. In our experiments, we randomly generated 10 different splits for each dataset, having 50 percent of people for training and the other 50 percent for testing. Then the results were averaged over the splits for evaluation. For each training, we initialized the model parameters randomly. We set the margin m in the verification loss to 2 and the dimensionality of embedded feature space d (the final output of Spatial ConvNet) to 256. The number of convolution kernels k for Temporal ConvNet was set to 32. We trained the whole network for 2,000 epochs with the learning rate set to 1e-6. During Training, we use the same 16 consecutive frames for each sequence as RCN did. During testing, given each probe sequence, we computed its deep features using trained network and used Euclidean distance for the result ranking. Data augmentation (cropping and flipping for each cropped image) was applied to both training and testing as done in the RCN work. In training the cropping was done at all 8 positions along the diagonal line with per-pixel shifting, while in testing we only cropped at the 4 positions of closest to the center of the original image. The results were averaged over all augmentation conditions.

Effectiveness and superiority of temporal convolution network

The biggest difference between our proposed T-CN model and the most related work RCN is that T-CN has a temporal convolution network, while RCN does not have. For a fair comparison, we reimplemented RCN and let it share exactly the same codes (including the same parameters) with T-CN for all the parts except temporal convolution network, which only exists in T-CN. The same training/test data splits are used. Following RCN, we use both color and optical flow for T-CN so that its temporal convolution network can focus more on the mid-level representation of motion, while optical flow captures low-level motion information.

The results in Figure 4 show that T-CN significantly outperforms RCN on both datasets, with a margin of 3.7% for PRID 2011 and 9.2% for iLIDS-VID in terms of Rank 1 accuracy. This indicates that temporal convolution is able to extract additional motion information that optical flow, RNN and temporal pooling cannot cover, which is likely to be the mid-level representation.

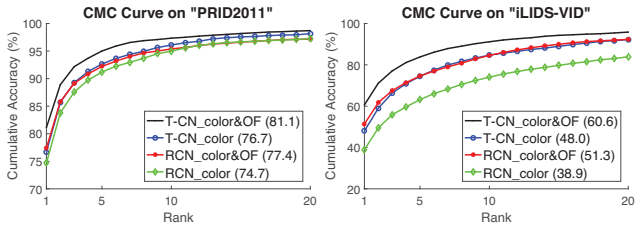

Figure 4: Effectiveness of temporal convolution network. The commonly used Cumulative Matching Characteristic (CMC) curves are adopted for evaluation. "Color" and "OF" indicate the color input and optical flow input, respectively. The Rank 1 accuracy (%) is shown in the legend as well.

Actually, T-CN is designed to be able to extract low-level motion information by itself. While we observe the same superiority of T-CN against RCN when color is the only input, it is interesting to see that T-CN with color only can get about the same performance as that of RCN with both color and optical flow. Note that for the result shown here, we had the temporal kernel size $S = 2$ which only models motion between adjacent frames just like what optical flow does. so it is reasonable to say that temporal convolution can actually replace the role of optical flow on low-level motion representation. This is very important, as it makes T-CN a pure end-to-end model, unlike RCN which relies on the non-learning based preprocessing of optical flow computation whose performance is algorithm dependent.

It is worth mentioning that compared with RCN, the additional costs for training and testing T-CN are ignorable. The model size (total number of parameters) of RCN is 827,424,

while that of T-CN (with $S = 2$) is 827,553, with only about 0.016% increment. According to our observation, the training time increased no more than 10%. Applying the two models for feature extraction at the testing stage takes about the same amount of time: for a sequence with 16 frames, RCN takes 0.3941 seconds while T-CN takes 0.3979 seconds with the same Nvidia Titan X GPU ($\sim 1\%$ difference).

Comparison with the state-of-the-art

To show the superiority of T-CN, we compare it with the state-of-the-art approaches for video-based person re-identification on the same datasets. The comparisons are given in Table 1 and 2. It is clear that our model shows superior performance against the competitors, especially when the most important Rank 1 accuracy is concerned.

Table 1: Comparison of T-CN with state-of-the-art methods on the PRID 2011 dataset, in terms of Cumulative Accuracy (%) at representative CMC ranks. The method with * is our implementation, while others are the original version from their authors. The lower half of the results (under the separation line) were got from a subset (89%) of the dataset with only the sequences longer than 21 frames. The numbers are shown in italic for differentiation, as that subset might make the task slightly easier. Results are round to 1 decimal place when applicable. The best result is shown in bold.

Methods	CMC Rank				Publication
	1	5	10	20	
T-CN (Ours)	**81.1**	**95.0**	**97.3**	98.7	This paper
RCN* (McLaughlin et al.'s)	77.4	92.2	95.4	97.2	CVPR 2016
CNN+XQDA (Zheng et al.'s)	77.3	93.5	N/A	**99.3**	ECCV 2016
RCN (McLaughlin et al.'s)	70	90	95	97	CVPR 2016
PaMM (Cho et al.'s)	45.0	72.0	85.0	92.5	CVPR 2016
AFDA (Li et al.'s)	43.0	72.7	84.6	91.9	BMVC 2015
DVDL (Karanam et al.'s)	40.6	69.7	77.8	85.6	ICCV 2015
SI^2DL (Zhu et al.'s)	***76.7***	***95.6***	*96.7*	***98.9***	IJCAI 2016
AvgTAPR (Gao et al.'s)	*68.6*	*94.6*	***97.4***	***98.9***	ICIP 2016
STFV3D+KISSME (Liu et al.'s)	*64.1*	*87.3*	*89.9*	*92.0*	ICCV 2015
TDL (You et al.'s)	*56.7*	*80.0*	*87.6*	*93.6*	CVPR 2016
DVR (Wang et al.'s)	*39.5*	*61.1*	*71.7*	*81.0*	TPAMI 2016

Kernel size of temporal convolution

The kernel size S of the temporal convolution is a key parameter of T-CN as it directly controls the scale/level of motion it learns to describe. Kernels of small sizes like $S = 2$ are more likely better at extracting low-level motion information, while those of larger sizes can probably cover mid-level motion description. Figure 5 show the results of T-CN with S ranging among 2,4,6 and 8 on the PRID 2011 dataset. It can be seen that smaller sizes generally perform better, and the slightly larger size $S = 4$ is slightly better than the smallest. Therefore, it is better to have T-CN cover the local motion (low-level and/or mid-level) beyond that happens between two adjacent frames only (what optical flow models).

Integration of kernel responses

A set of kernels have been used for temporal convolution to describe different motion patterns. Instead of simpling pooling the results in the normal way such as using average pooling (maximum pooling does not fit here as we expect to integrate information from all kernels), we propose applying a full-length 1D convolution for a weighted averaging where the weights (i.e., parameters of the convolution kernel) can be automatically learned from the training data. Table 3 compares the results of these two strategies. Obviously, though being very light, learned weighting in terms of full-length 1D convolution is significantly better than the simple average pooling. One may also use a dense layer to do the integration, but it did not show good results in our own experiments.

Table 2: Comparison of T-CN with state-of-the-art methods on the iLIDS-VID dataset. The method with * is our implementation, while others are the original version from their authors. The best result is shown in bold.

Methods	CMC Rank				Publication
	1	5	10	20	
T-CN (Ours)	**60.6**	83.8	91.2	95.8	This paper
RCN (McLaughlin et al.'s)	58	84	91	96	CVPR 2016
TDL (You et al.'s)	56.3	**87.6**	**95.6**	**98.3**	CVPR 2016
AvgTAPR (Gao et al.'s)	55.0	87.5	93.8	97.2	ICIP 2016
RCN* (McLaughlin et al.'s)	53.3	77.2	86.8	93.8	CVPR 2016
CNN+XQDA (Zheng et al.'s)	53.0	81.4	N/A	95.1	ECCV 2016
SI^2DL (Zhu et al.'s)	48.7	81.1	89.2	97.3	IJCAI 2016
STFV3D+KISSME (Liu et al.'s)	44.3	71.7	83.7	91.7	ICCV 2015
DVR (Wang et al.'s)	39.5	61.1	71.7	81.0	TPAMI 2016
AFDA (Li et al.'s)	37.5	62.7	73.0	81.8	BMVC 2015
PaMM (Cho et al.'s)	30.3	56.3	70.3	82.7	CVPR 2016
DVDL (Karanam et al.'s)	25.9	48.2	57.3	68.9	ICCV 2015

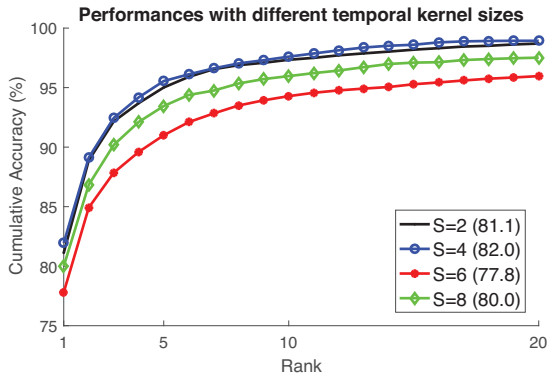

Figure 5: Performances of T-CN on PRID 2011 dataset when different kernel sizes S are used for temporal convolution.

Table 3: Comparison of different pooling methods for temporal convolution filter response integration, in terms of Cumulative Accuracy (%) at representative CMC ranks.

Dataset	PRID 2011				iLIDS-VID			
CMC Rank	1	5	10	20	1	5	10	20
Learned Weighting	81.1	95.0	97.3	98.7	60.6	83.8	91.2	95.8
Average Pooling	75.1	91.6	95.8	97.9	57.5	81.4	89.4	95.0

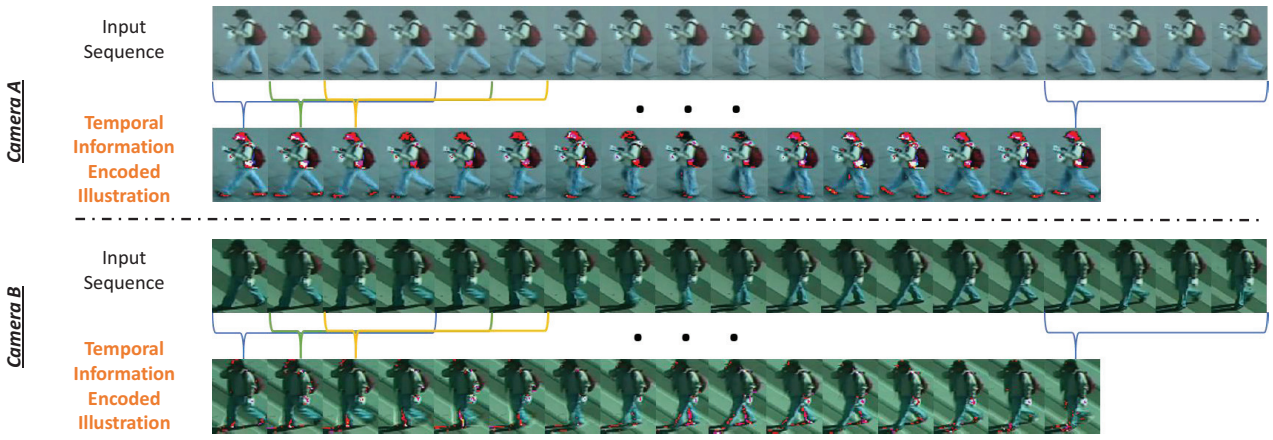

Figure 6: Visualizing the illustrative effect of temporal convolution. Please refer to the text for details.

Cross-Dataset Testing

A very important expectation for any person re-identification model towards real-world applications is the ability to generalize over different camera networks, as it is undesirable or even impossible to provide high quality labeled training data (which is a hard task for humans) for each network before application. Cross-dataset testing is a simple way to test such generalization ability. We follow the same setting as the RCN work (McLaughlin, Martinez del Rincon, and Miller 2016) and compare with it, as it got significantly superior results than other existing work.

Table 4 reports our results on both datasets. Though the performances show that cross-dataset generalization is very challenging and the state-of-the-art is still far from being satisfactory, our T-CN can do a better job on it. This is encouraging, as it shows the learned low-level and middle-level motion representation is more transferable than what optical flow brings. We expect more work can be done later to improve such a generalization ability.

Table 4: Cross-dataset testing performance comparison on PRID 2011 and iLIDS-VID, respectively (trained on the other dataset), in terms of Cumulative Accuracy (%) at representative CMC ranks. RCN* is our implementation which is made most comparable to our T-CN model.

Test Dataset	**PRID 2011**				**iLIDS-VID**			
CMC Rank	1	5	10	20	1	5	10	20
T-CN	24.0	50.5	64.2	78.5	13.4	30.2	40.8	54.6
RCN* (CVPR 2016)	25.0	47.0	57.9	71.9	11.0	27.1	38.8	52.5

Visualizing the effect of temporal convolution

To understand why T-CN is able to learn low-level and middle-level motion representation, we visualize the illustrative effect of temporal convolution in Figure 6 using a concrete example from our experiment on the PRID 2011 dataset. From a trained model with kernel size $S = 4$, we choose the temporal convolution kernel which gets the largest learned weight during kernel response integration, and use it to convolute along the time axis the input RGB video sequences (one sequence from each camera) of an arbitrary person. Then for each input sequence the temporal convolution will generate a new sequence with each of its elementary image being able to encode information from 4 continuous frames from the original input sequence, as shown in Figure 6. For a better illustrative visualization, we enhance each output image by adding to this output image the difference image between it and the first input frame of the 4 continuous frames that the kernel covers. By doing so, the information related to motion gets more visible. From the figure, we can see that the output is able to encode both the appearance information and the motion information (in terms of the overlapping shape patterns). Interestingly, the enhancement makes some parts turn red, which seem to be positions with relatively larger motion in each image.

Conclusion and future work

In this paper we introduce a new network architecture called T-CN for video-based person re-identification. T-CN contains a temporal convolution network being able to learn low-level and mid-level motion representation which was hard to achieve by existing solutions. T-CN allows building a whole end-to-end learnable representation model with recurrent neural network and temporal pooling, which can process video sequences of various lengths in an efficient way. Experiments on commonly used benchmark datasets show the superiority of our proposal to the state-of-the-art, in terms of both effectiveness and generalization ability.

In the future, we'd like to extend and evaluate T-CN on recently released larger dataset Mars (Zheng et al. 2016) with tracking noises. Even a naive application of the simplest T-CN with many limitations: very shallow Spatial ConvNet), no optical flow, no data augmentation, no parameter tuning, single query, and no re-ranking, already shows encouraging results (rank1 accuracy of 61.2%), comparable to the state-of-the-art (Zheng et al. 2016; Zhong et al. 2017). We believe a much better performance can be got with deeper Spatial ConvNet (e.g. ResNet) and other improvements.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 15K16024 and 16K12421 and the National Natural Science Foundation of China Under Grant No.61373077. The corresponding author is Yang Wu.

References

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2360 – 2367.

Gao, C.; Wang, J.; Liu, L.; Yu, J.; and Sang, N. 2016. Temporally aligned pooling representation for video-based person re-identification. In *IEEE International Conference on Image Processing (ICIP)*, 4284–4288.

Gray, D., and Tao, H. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 262–275.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, 1735–1742.

Hamdoun, O.; Moutarde, F.; Stanciulescu, B.; and Steux, B. 2008. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *2008 Second ACM/IEEE International Conference on Distributed Smart Cameras*, 1–6.

Han, J., and Bhanu, B. 2006. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(2):316–322.

Hirzer, M.; Beleznai, C.; Roth, P. M.; and Bischof, H. 2011. Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*.

Karanam, S.; Li, Y.; and Radke, R. J. 2015. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*.

Köstinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P. M.; and Bischof, H. 2012. Large scale metric learning from equivalence constraints. In *CVPR*, 2288–2295.

Layne, R.; Hospedales, T. M.; Gong, S.; and Mary, Q. 2012. Person re-identification by attributes. In *BMVC*, volume 2, 8.

Li, Y.; Wu, Z.; Karanam, S.; and Radke, R. J. 2015. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *Proceedings of the British Machine Vision Conference (BMVC) 2015, Swansea, UK, September 7-10, 2015*, 73.1–73.12.

Li, D.; Zhang, Z.; Chen, X.; Ling, H.; and Huang, K. 2016. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*.

Liao, S.; Hu, Y.; Zhu, X.; and Li, S. Z. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*.

Liu, K.; Ma, B.; Zhang, W.; and Huang, R. 2015. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *ICCV*.

Matsukawa, T.; Okabe, T.; Suzuki, E.; and Sato, Y. 2016. Hierarchical gaussian descriptor for person re-identification. In *CVPR*.

McLaughlin, N.; Martinez del Rincon, J.; and Miller, P. 2016. Recurrent convolutional network for video-based person re-identification. In *CVPR*.

Mignon, A., and Jurie, F. 2012. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2666–2672.

Pedagadi, S.; Orwell, J.; Velastin, S.; and Boghossian, B. 2013. Local fisher discriminant analysis for pedestrian re-identification. In *CVPR*.

Shi, Z.; Hospedales, T. M.; and Xiang, T. 2015. Transferring a semantic representation for person re-identification and search. In *CVPR*.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *ECCV*, 688–703.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2016. Person re-identification by discriminative selection in video ranking. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(12):2501–2514.

Wu, Y.; Minoh, M.; Mukunoki, M.; and Lao, S. 2012. Set based discriminative ranking for recognition. In Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; and Schmid, C., eds., *ECCV*, volume 7574 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 497–510.

Wu, Y.; Li, W.; Mukunoki, M.; Minoh, M.; and Lao, S. 2015. Discriminative collaborative representation for classification. In Cremers, D.; Reid, I.; Saito, H.; and Yang, M.-H., eds., *Computer Vision – ACCV 2014*, volume 9006 of *Lecture Notes in Computer Science*. Springer International Publishing. 205–221.

You, J.; Wu, A.; Li, X.; and Zheng, W.-S. 2016. Top-push video-based person re-identification. In *CVPR*.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. Mars: A video benchmark for large-scale person re-identification. In *ECCV*. Springer.

Zheng, L.; Yang, Y.; and Hauptmann, A. G. 2016. Person re-identification: Past, present and future. *CoRR* abs/1610.02984.

Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*.

Zhu, X.; Jing, X.; Wu, F.; and Feng, H. 2016. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *IJCAI*, 3552–3559.