Image-to-Video Person Re-Identification With Temporally Memorized Similarity Learning

Dongyu Zhang, Wenxi Wu, Hui Cheng, Ruimao Zhang, Zhenjiang Dong, and Zhaoquan Cai

***Abstract*— With the development of video surveillance in public safety field, there is an increasing research on person re-identification (re-id). In this paper, we address the image-to-video person re-id, in which the probe is an image and the gallery is consists of videos captured by nonoverlapping cameras. Compared with image, video sequence contains more temporal information that can be explored to improve the performance of re-identification system. However, it is challenging to model temporal information in the matching process of image-to-video person re-id. In this paper, we proposed a novel temporally memorized similarity learning neural network for this problem. In specific, the proposed network mainly consisted of two parts, including feature representation sub-network and similarity sub-network. In the first part, we adopted a convolutional neural network (CNN) to extract features from the input image. Given a video sequence of a person, features were first extracted from each its frame by using CNN and further forward to a long shot term memory (LSTM) network to encode the temporal information of video sequence. The outputs of LSTM were concatenated together as the feature vector of video sequences. Finally, the feature vectors of probe image and the video sequence were further forward to the similarity sub-network for distance metric learning. In the proposed framework, the feature representation and the similarity metric learning can be learned and optimized simultaneously. We evaluated the proposed framework on three public person re-id data sets, and the experimental results showed that the proposed approach is effective for the image-to-video person re-id.**

***Index Terms*— Person re-id, deep metric learning, LSTM.**

I. Introduction

PERSON re-identification (re-id) refers to the problem of recognizing people across images and videos from non-overlapping camera views. It has attracted increasing interests due to its broad applications in video surveillance, such as people tracking [1] and multi-person association [2]. Despite the best efforts of many researchers of decades years, person re-id remains a very challenging problem. The main reason is that a person's appearance often changes dramatically across camera views due to the large variations in illumination, poses, viewpoints and background images and videos [3]–[5].

Most of existing literature of person re-id are based on the similarity matching between still images. Given a probe image of one person, the goal is to find images of the person with the same ID in the gallery based on their similarity. A gallery is consists of images cropped from a different view of videos of unknown persons. However, in the still image-based person re-id, the temporal information of video sequences is often ignored, which limits the recognizing performance, especially when the person is occluded by objects or other persons.

Benefiting from the more and more surveillance cameras installed in public place, it is becoming more convenient to acquire videos of a pedestrian. Thus, identifying a person directly by using the videos has received more attentions recently. For example, given an image of an escaped criminal, the policeman wants to find out the escape route by using the surveillance videos. In this case, the matching is between image and videos. We call this kind of person re-id to be image-to-video person re-id, where the probe is an image and the gallery is consisting of videos captured by nonoverlapping cameras.

Actually, the use of video sequence is a natural way for person re-id. As videos are collections of images arranged according to the timestamps, temporal information related a person can be captured to deal with difficult cases of recognizing a person from different cameras. Furthermore, a sequence of images provides numerous samples of a person's appearance, where the samples may have a different pose, viewpoint as well as the background. Thus a better model of the person's appearance can be built. Besides, from the point view of the application, the image-to-video person re-id is more convenient to still image-based person re-id, as there is no need to crop pedestrians from the videos to construct the gallery.

Although more information can be obtained from videos, image-to-video person re-id remains a challenging problem. Like in the still-image-based method, the problem of people with similar appearance have similar representations still remains in image-to-video person re-id. Although person's motions is an important behavioral biometrics cue for identifying different persons, it is unfortunate that the walking actions of them in videos may be similar as well. As pointed out by [6] that for some instances, it is harder to distinguish the video representations of different identities than the still image cases. Besides, as the similarity matching is between two different

Manuscript received January 30, 2017; revised June 4, 2017; accepted June 25, 2017. Date of publication July 4, 2017; date of current version October 24, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61401125, Grant 61671182, Grant U1611461 and Grant 61370185; in part by the NSFC-Shenzhen Robotics Projects under Grant U1613211; in part by Guangdong Natural Science Foundation under Grant 1614050001452; in part by the Fundamental Research Funds for the Central Universities; and in part by ZTE Corporation. This paper was recommended by Associate Editor W. Zuo. *(Corresponding author: Hui Cheng.)*

D. Zhang, W. Wu, and H. Cheng are with Sun Yat-sen University, Guangzhou 510006, China (e-mail: chengh9@mail.sysu.edu.cn).

R. Zhang is with Chinese University of Hong Kong, Hong Kong 999077, China.

Z. Dong is with ZTE Corporation, Shenzhen 518057, China.

Z. Cai is with Huizhou University, Huizhou 516007, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2017.2723429

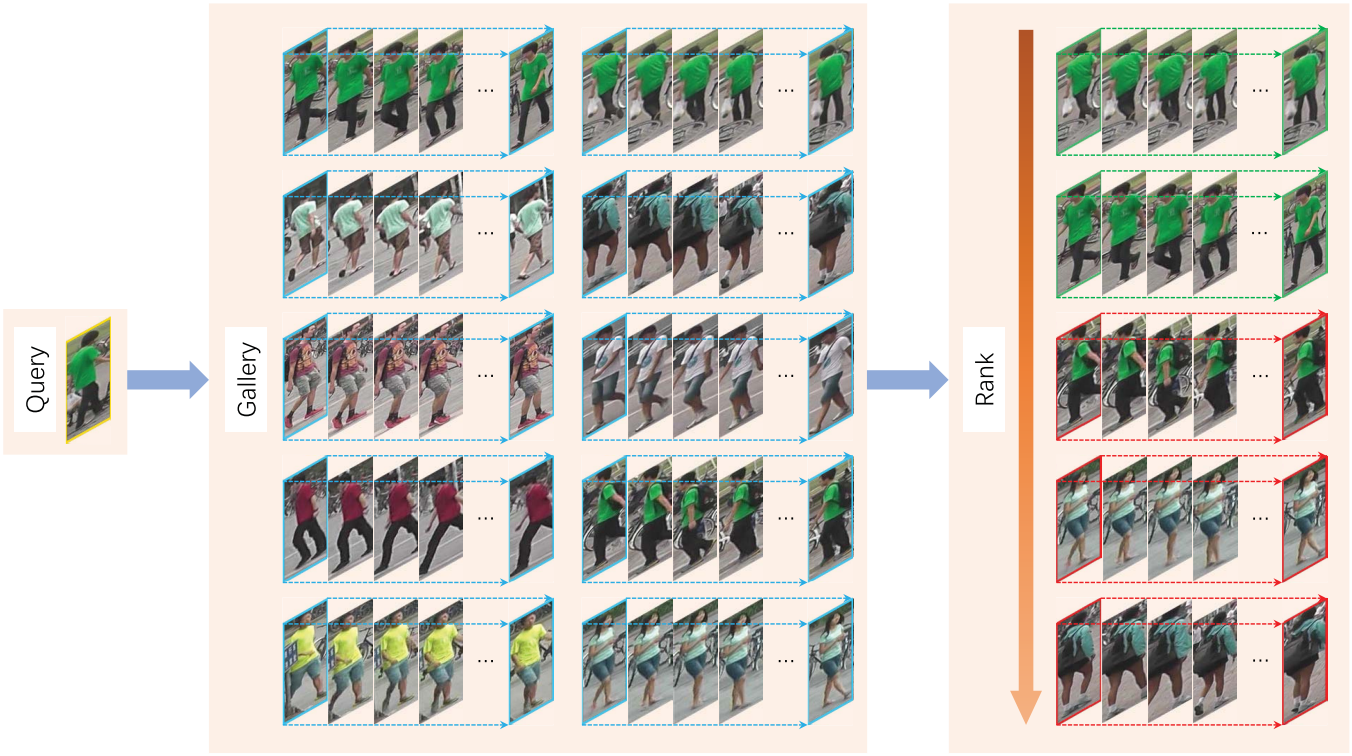

Fig. 1. Illustration of image-to-video person re-id problem. In image-to-video person re-id, the probe is an image, while the gallery is consists of different viewpoint videos of pedestrians. The aim of this problem is to find videos containing the same person id with the probe image, and ranking the videos in gallery according to their similarity to the probe image.

modalities, i.e., image and video, how to effectively represent their features and measure their similarity is also crucial for the accuracy of the re-id system.

In order to solve the above problems, we propose a novel neural network framework for image-to-video person re-id in this paper. The proposed framework formulates the feature extraction, video spatial-temporal information encoding, and similarity learning in an end-to-end way. Concretely speaking, the proposed framework consists of a convolutional neural network (CNN), a long shot term memory (LSTM) network, and a similarity sub-network. During the training step, we adopt CNN for the dense feature extraction of an image of a person. For the videos sequence of the same person, the features are extracted from each frame of videos using CNN, which is incorporated into an LSTM network to further encode the spatial and temporal information in videos. As a recurrent neural network, LSTM allows information to flow between time-steps of the video sequence. The outputs of LSTM at each time-steps are concatenated together as the features of videos. And finally, the features of image and videos are forward to the similarity sub-network to learn a distance metric to measure the similarity between the image and videos. The architecture of the proposed framework is shown in Fig. 1.

The key contribution of this paper is that we proposed a general cross-modality matching framework for image-to-video person re-identification, which adopts CNN and LSTM network for the deep feature extraction and temporal information of video encoding, and further adopts a neural network for similarity measure learning. Our method can directly learn both spatial features and temporal feature from a video sequence in an end-to-end manner. By using this framework, the feature representation and similarity measure can be learned and optimized simultaneously. We perform extensive experiments on three public person re-id datasets to evaluate the proposed methods. Experimental results demonstrate that the proposed method is effective for the image-to-video person re-id.

The remainder of this paper is organized as follows. In Section II, we briefly review the related literature about person re-id in recent years. Section III describes the proposed framework and presents its each part in detail. Extensive experimental results and discussion are provided in Section IV. Finally, Section V concludes this paper.

II. Related Work

In recent years, person re-id has attracted increasing interests due to its broad application. According to the way of feature representation, we divide the existing literature of person re-id into two categories, i.e., hand-crafted feature-based person re-id, and deep learning-based person re-id. In the following, we give a brief review on the two kinds of person re-id methods, respectively.

A. Hand-Crafted Feature-Based Person Re-Id

To the best of our knowledge, since the work of Gheissari *et al.* [7] in 2006, the hand-crafted feature of a single image has been explored for person re-id. Generally,

there are two main components for hand-crafted feature-based person re-id system, i.e., feature representation and distance metric learning. Features are first extracted from the probe image and gallery images, and then certain distance metric is learned to measure the similarity of these features across different images. Therefore, the research of these methods usually focuses either on feature representation to construct discriminative features or on distance metric learning to find an improved similarity metric for feature matching.

1) Feature Representation: Commonly the most popular features used for hand-crafted feature-based person re-id tasks are color and textures, such color histograms, local binary patterns (LBP), local maximal occurrence (LOMO) and their variants. For example, in the work of [7] Gheissari *et al.* compute the HS histogram and edge histogram for the pedestrian description for person re-id. Gray and Tao [8] partition the pedestrian into horizontal stripes, and use 8 color channels, i.e., RGB, HS, and YCbCr for pedestrian description. Similarly, Mignon and Jurie [9] build the feature vector from RGB, YUV and HSV channels and extract the LBP texture histograms in horizontal stripes. Zhao *et al.* adopt a 32-dim LAB color histogram and a 128-dim SIFT descriptor to extract features from each patch of 10×10 of the pedestrian. Li *et al.* [10] extract local color descriptors from patches and further using hierarchical Gaussianization [11] to capture spatial information. In [12] color histograms and moments are extracted from HSV and YUV spaces. Zuo *et al.* [13] proposed a generalized shrinkage-thresholding operators for blind deconvolution. Liu *et al.* [14] extract the HSV histogram, gradient histogram and the LBP histogram for each local patch. Yang *et al.* [15] introduce the salient color names based color descriptor for pedestrian color descriptions. Lin *et al.* [16] proposed and-or graph models for object shape detection. Liao *et al*. [17] propose the LOMO descriptor, which includes the color and SILTP histograms. In [18], Zheng *et al.* propose extracting the 11-dim color descriptor for each local patch.

2) Distance Metric Learning: Apart from the discriminative hand-crafted features designing, distance metric learning is another important component for handcraft feature-based person re-id. Most of the current metric learning method of person re-id tends to learn a Mahalanobis distance functions which enlarge the distance of mismatched images and reduces that of matched images. For example, Weinberger and Saul [4] proposed the large margin near neighbor learning (LMNN) model, in which the distance metric was learned to separate the matched neighbors from the mismatched ones by a large margin. Davis *et al.* [19] further improved the overfitting problem of this model and proposed the information-theoretic metric learning (ITML) which made a trade-off between satisfying the given similarity constraints and ensuring that the learned metric to be close to the initial distance function. Jing *et al.* [20] proposed a super-resolution person re-id method with semi-coupled low-rank discriminant dictionary learning. Cheng *et al.* [21] proposed to combine the bilinear similarity with the Mahalanobis distance to model the cross-patch similarities. Guillaumin *et al.* proposed a logistic discriminant metric learning (LDML) model by modeling the probability of a given sample pair (x_i, x_j) and used the maximum log-likelihood as the objective function [22]. Koestinger *et al.* [23] proposed a KISSME learning method to address the scalability issue of metric learning from equivalence constraints. Liao *et al.* [17] improved the KISSME method by learning a discriminant low dimensional subspace based on the LOMO features. They also improved the LDML model by enforcing the positive semidefinite constraint and the asymmetric sample weighting strategy [24].

B. Deep Learning-Based Person Re-ID

In recent years, deep learning models have achieved great success in different kind of tasks in computer vision, such image classification [25], objection detection [26], image denoising [27] and etc. [28]–[33]. Due to its power in learning discriminative features from large-scale image data, many methods have adopted deep learning models to jointly learn the representation and the classifier [34]–[36]. The first two works of deep learning-based person re-id were [37], [38]. Currently, the most commonly used model in deep learning-based person re-id is Siamese model. For example, Li *et al.* [38] added a patch matching layer to Siamese model to multiplies the convolutional responses of two images in different horizontal stripes. Ahmed *et al.* [34] improved the Siamese model by computing the input neighborhood difference features to compares the features from neighboring locations of the cross image.

Usually in Siamese network ranking functions were learned based on pairs [37] or triplets [39] of images. Schroff *et al.* adopted a deep CNN to learn the Euclidean embedding per image by using the triplet comparison loss [40]. Cheng *et al.* [21] designed a triplet loss function that takes three images as input. Ding *et al.* proposed a deep learning model based on relative distance comparison for person re-id [41]. Su *et al.* [42] proposed a three-stage learning process which includes attribute prediction using an independent dataset and an attributes triplet loss trained on datasets with ID labels. Varior *et al.* [43] proposed to insert a gating function after each convolutional layer to capture effective subtle patterns when a pair of testing images were fed into the network. Liu *et al.* [44] proposed integrating a soft attention-based model in a Siamese network to adaptively focus on the important local parts of an input image pair. Wang *et al.* [31] proposed a joint learning deep CNN framework, in which the matching of single-image representation and the classification of cross image representation are jointly optimized for pursuing better matching accuracy.

Currently, most of the methods of person re-id are based on image-to-image matching, either single-shot matching or multi-shot matching. With the uprising application of intelligent video surveillance, video-based person re-id has attracted more and more attentions. Simonnet *et al.* [45] proposed to use dynamic time warping to solve the sequence matching problem in video-based person re-id. Zheng *et al.* [46] proposed to train a CNN classification model for video-based person re-id by using the frames of an identity as its training samples. Fernando et al. proposed

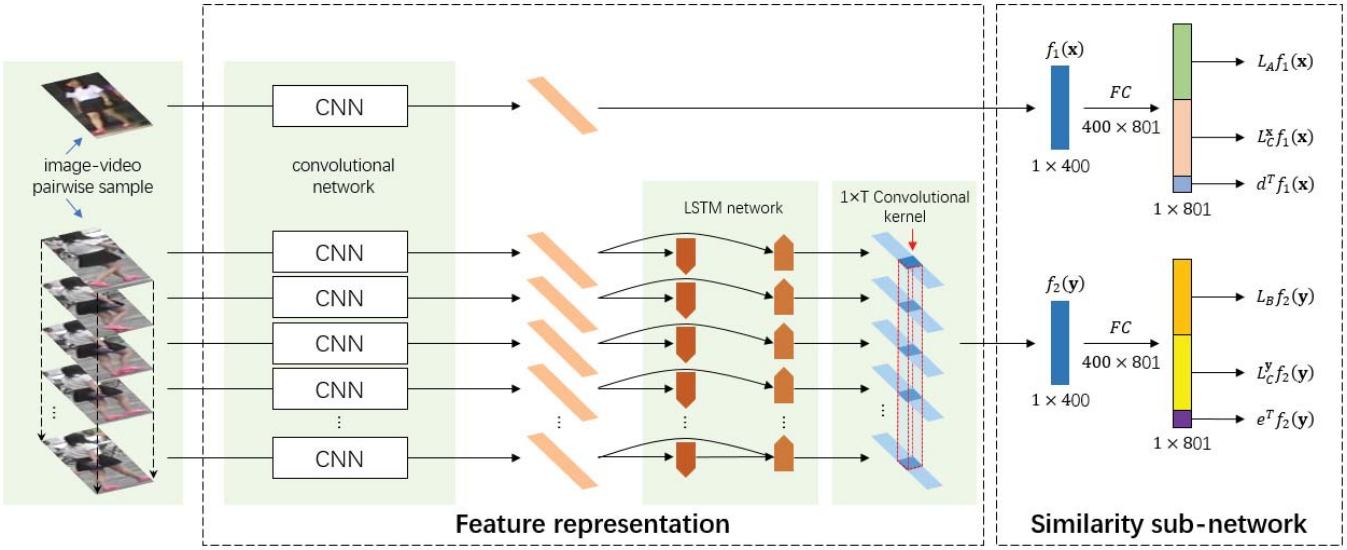

Fig. 2. The pipeline of proposed framework. There are mainly two parts in the proposed framework, including feature representation sub-network, and similarity sub-network. In the feature representation sub-network, the feature vector of the input image is extracted by CNN. For a video, CNNs are used to extract features each of its frame and LSTMs are adopted to further encode the temporal information of video sequence. Finally, similarity sub-network is used for distance metric learning.

a learning-to-rank model to capture the video frame features. Wang *et al.* [47] embedded a multi-level encoding layer into the CNN model and produced video descriptors of varying sequence lengths.

Some works also use the recurrent neural network for the video-based person re-id. For example, in [48] McLaughlin *et al.* proposed to use RNN to learn the interaction between multiple frames in a video. In their work, they made use of color and optical flow information to capture appearance and motion information, and they adopt a temporal pooling layer in the RNN to address the problem that the output tends to bias toward later time-steps. Their network was joint trained for the identification and verification. In [49], to address the multi-shot person re-id, Yan *et al.* proposed a recurrent feature aggregation network (RFA-Net) that builds sequence level representation from a temporally ordered sequence of frame-wise feature based on RNN network. LBP and color features of an image were the first extract and concatenated as frame level representation, which was forward to RNN network for further processing. Finally, RankSVM was adopted to measure the similarity of features. Our work was significantly different from their works in that we fused the feature representation and the distance metric learning in the same framework. Thus, we can jointly learn the feature representation and distance metric and perform the optimization simultaneously, which was helpful in improving the re-id performance. Besides, in this paper, we addressed the problem of image-to-video person re-id, where the probe was an image and the gallery consists of video sequences of different viewpoints.

III. Methodology

In this paper, we propose a new framework to address the image-to-video person re-id problem, which is different from the single image-based person re-id in that the gallery is not consists of images of pedestrians but their video tracklets with arbitrary length. Besides, as the similarity is compared between image and video in image-to-video person re-id, how to represent the features of videos and perform the cross-modality similarity matching are crucial to the re-id accuracy.

CNN models have been proved to be effective in image classification task, and several approaches have been proposed to apply CNNs to person re-id [34], [38]. Nonetheless, these CNN-based architectures are dealing with single-shot image, and are not directly applicable to image-to-video person re-id as the videos tracklets often have arbitrary length.

Recently, recurrent neural networks, particularly LSTMs, have been proved to be stable and powerful in modeling long-range dependencies in learning tasks. Thus, in this work, we propose to combine CNN models and LSTM network for the feature representation of image and video. In addition, we adopt a similarity sub-network to handle the cross-modality matching problem in image-to-video person re-id. In this section, we first give an overview of the proposed framework, and then describe each component in detail.

A. Architecture Overview

Fig. 2 shows the architecture of the proposed framework. Generally, there are mainly two parts, i.e., feature representation sub-network and similarity sub-network in the proposed framework. The first part is used for image and video feature extraction, while the second part is used for similarity learning. In the first part, we use a CNN to extract the feature of an input image and adopt the combination of CNNs and LSTM to extract the feature of videos. Each frame of videos is first processed by CNN to produce a feature vector representing the persons appearance, which is forward to the LSTM to further exploit spatial and temporal information within

the video sequence. The outputs of LSTM are concatenated together as the feature vector of the video sequence. Finally, the features of the input image and video are forward to the similarity sub-network for distance metric learning.

By using the proposed framework, we can conveniently extract the features of image and video, and calculate their similarity. Besides, as the similarity learning is embedded in the neural network, we can optimize the feature representation and similarity learning simultaneously.

B. Feature Representation Sub-Network

As shown in Fig. 2, we deploy the feature representation sub-network to extract the features of the input image and video. It is commonly accepted that the performance of deep networks is due to hierarchical feature extraction that takes place over many layers. Therefore for the input image, we adopt a CNN to extract its features. As to the input video, we use the CNNs with the same architecture to pre-process each frame into a higher-level representation and then adopt an LSTM network to further explore the spatial and temporal information in the video sequence.

1) Convolutional Network: By using a hierarchy of trainable filters and feature pooling operations, CNN is capable of automatically learning complex features for understanding image content, and achieving superior performance to hand-crafted features [34]. Encouraged by these positive results, we adopt the CNN for feature extraction in our framework. We refer to CNN as a function $C(\cdot)$. Given an input image x, CNN outputs a $d \times 1$ dimension vectorized representation of its final layer activation maps as the feature vector of x,

$$\mathbf{f}(x) = C(x). \tag{1}$$

For the input video, we also use CNNs with the same architecture to extract the features of its frames. Let $\mathbf{y} = \{y^{(t)} | t \in [1, \ldots, T]\}$ be a video sequence, where $y^{(t)}$ is the frame at time t consisting of whole-body images of a person and T is the length of video sequence. Each frame of $\mathbf{y}$ is passed through the CNNs separately to produce a vector as,

$$f_{\mathbf{y}}^{(t)} = C(y^{(t)}), \tag{2}$$

where $t \in [1, \ldots, T]$. Each vector $f_{\mathbf{y}}^{(t)}$ is then passed forward to the LSTM layer, where it is projected into a low-dimensional feature space and combined with information from previous time steps. Note that the parameters of the CNNs are shared across all time-steps meaning that each input frame is processed by the same feature extraction network.

2) LSTM Network: After pre-processing by CNN, the extracted features of video are forward to the LSTM layers. As image-to-video person re-id involves recognizing a person from a video containing a time-series of images, the use of LSTM layers may help to improve person re-id performance by allowing information to be passed between time-steps. By incorporating LSTM layers, we aim to better capture temporal information present in the video. At each time-step, the LSTM receives a new input and produces an output based on both the current input and information from the previous time-steps. During the training of LSTM network by using back-propagation, the recurrent connections are unrolled to create a very deep feed-forward network [50]. The LSTM network can easily memorizes the long-period interdependencies in sequential data, and the temporal dependency learning can be conveniently converted to the spatial domain.

In our framework, each LSTM accepts the previous single video frame feature vector $f_{\mathbf{y}}^{(t)}$ in Eq.(2) as input and determines the current states that comprises the hidden cells $\mathbf{h}_{\mathbf{y}}^{t+1} \in \mathbb{R}^d$ and the memory cells $\mathbf{m}_{\mathbf{y}}^{t+1} \in \mathbb{R}^d$, where d is the output number. Following [51], in our work the LSTM network consists of four gates: the input gate g^u, the forget gate g^f, the memory gate g^c and the output gate g^o. The W^u, W^f, W^c, W^o are the corresponding recurrent gate weight matrices. Suppose, $\mathbf{H}_{\mathbf{y}}^t$ is the concatenation of the input $f_{\mathbf{y}}^{(t)}$ and the previous states is $\mathbf{h}_{\mathbf{y}}^t$. The hidden and memory cells can be updated as follows:

$$\begin{aligned} g^u &= \sigma(W^u * \mathbf{H}_{\mathbf{y}}^t), \\ g^f &= \sigma(W^f * \mathbf{H}_{\mathbf{y}}^t), \\ g^o &= \sigma(W^o * \mathbf{H}_{\mathbf{y}}^t), \\ g^c &= tanh(W^c * \mathbf{H}_{\mathbf{y}}^t), \\ \mathbf{m}_{\mathbf{y}}^{t+1} &= g^f \odot \mathbf{m}_{\mathbf{y}}^t + g^u \odot g^c \\ \mathbf{h}_{\mathbf{y}}^{t+1} &= \tanh(g^o \odot \mathbf{m}_{\mathbf{y}}^{t+1}) \end{aligned} \tag{3}$$

where σ is the logistic sigmoid function, and $\odot$ indicates a pointwise product. The memory cell $\mathbf{h}_{\mathbf{y}}^t$ of LSMT at each time t is concatenated together as the features of the video sequence.

Although LSTMs are able to accumulate the information of videos, a possible drawback is that the LSTM's output may be towards to the later time-steps, which could reduce the performance of LSTM when used to accumulate the input information over a full sequence. To overcome this drawback, McLaughlin *et al.* [48] used the average-pooling over the temporal dimension to produce a single feature vector. Let $\mathbf{h}_{ave}$ represents a person's appearance information over the whole input sequence. Given the temporal pooling layer inputs $\{\mathbf{h}_{\mathbf{y}}^1, \mathbf{h}_{\mathbf{y}}^2, \ldots, \mathbf{h}_{\mathbf{y}}^T\}$, the average-pooling method is as follows:

$$\mathbf{h}_{ave} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_{\mathbf{y}}^t \tag{4}$$

where T is the length of the sequence. This average strategy can be regarded as to assign equal weight $\frac{1}{T}$ to output vector $\mathbf{h}_{\mathbf{y}}^t$ of LSTM at each time-step.

As we know that for person re-id, the features of some video frames may contain more discriminative information helpful in the re-id progress than others. Thus, those features should be assigned more weight. However, the average pooling strategy of Eq. (4) could not fully explore this information. In order to overcome this drawback, we apply a convolutional layer on the features of LSTM to adaptively learn the weights of each feature vector. Let $\mathbf{H}_{d \times T}$ denote the concatenated features map of LSTM as shown in Fig. 2, where $\mathbf{H}_{d \times T} = [\mathbf{h}_{\mathbf{y}}^1 \mathbf{h}_{\mathbf{y}}^2 \ldots \mathbf{h}_{\mathbf{y}}^T]$. After applying a $1 \times T$ convolutional kernel $W_{1 \times T}$ on $\mathbf{H}_{d \times T}$,

we can obtain a d dimension feature vector $\mathbf{f}(\mathbf{y})$ as

$$\mathbf{f}(\mathbf{y}) = \mathbf{H}_{d\times T} \otimes W_{1\times T}, \tag{5}$$

where $\otimes$ denotes the convolution operation. By using the Eq.(2) and Eq.(5), we can obtain the feature vector $\mathbf{f}(x)$ of input image x, and the feature vector $\mathbf{f}(\mathbf{y})$ of video $\mathbf{y}$ respectively. The two feature vectors are further forward to the similarity sub-network for the distance metric learning.

C. Similarity Sub-Network

Recently, Lin *et al.* [52] proposed a generalized similarity measure by fusing the affine Mahalanobis distance and Cosine similarity, which is further embedded into CNN for distance metric learning. Motivated by their work, we adopt a similarity sub-network in our framework to learn a distance metric to measure the similarity between the image and the video. In this case, we unify the feature representation and distance metric learning in the same framework, which jointly optimizes the two parts simultaneously.

Suppose that $\mathbf{f}(x) \in \mathbb{R}^d$, $\mathbf{f}(\mathbf{y}) \in \mathbb{R}^d$ are the feature of image and video, the generalize similarity measure of [52] is defined as:

$$S(\mathbf{f}(x), \mathbf{f}(\mathbf{y})) = \begin{bmatrix}\mathbf{f}(x)^{\mathbf{T}} & \mathbf{f}(\mathbf{y})^{\mathbf{T}} & 1\end{bmatrix} \begin{bmatrix}\mathbf{A} & \mathbf{C} & \mathbf{d} \\ \mathbf{C}^{\mathbf{T}} & \mathbf{B} & \mathbf{e} \\ \mathbf{d}^{\mathbf{T}} & \mathbf{e}^{\mathbf{T}} & \alpha\end{bmatrix} \begin{bmatrix}\mathbf{f}(x) \\ \mathbf{f}(\mathbf{y}) \\ 1\end{bmatrix} \tag{6}$$

where sub-matrices $\mathbf{A}$ and $\mathbf{B}$ are positive semi-definite, representing the self-correlations of the samples in their own domains, and sub-matrices $\mathbf{C}$ is a correlation matrix crossing the two domains. Parameter $\mathbf{d}, \mathbf{e}$ are vectors and α is a scalar. Thus, $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ can be further factorized as:

$$\begin{aligned}\mathbf{A} &= \mathbf{L}_{\mathbf{A}}^{\mathbf{T}}\mathbf{L}_{\mathbf{A}} \\ \mathbf{B} &= \mathbf{L}_{\mathbf{B}}^{\mathbf{T}}\mathbf{L}_{\mathbf{B}} \\ \mathbf{C} &= -\mathbf{L}_{\mathbf{C}}^{x\,\mathbf{T}}\mathbf{L}_{\mathbf{C}}^{\mathbf{y}}\end{aligned} \tag{7}$$

By using Eq.(7), the generalized similarity measure of Eq.(6) can be rewritten as following model:

$$\begin{aligned}S(\mathbf{f}(x), \mathbf{f}(\mathbf{y})) = &\|\mathbf{L}_{\mathbf{A}}\mathbf{f}(x)\|^2 + \|\mathbf{L}_{\mathbf{B}}\mathbf{f}(\mathbf{y})\|^2 + 2\mathbf{d}^{\mathbf{T}}\mathbf{f}(x) \\ &- 2(\mathbf{L}_{\mathbf{C}}^{x}\mathbf{f}(x))^{\mathbf{T}}(\mathbf{L}_{\mathbf{C}}^{\mathbf{y}}\mathbf{f}(\mathbf{y})) + 2\mathbf{e}^{\mathbf{T}}\mathbf{f}(\mathbf{y}) + \alpha\end{aligned} \tag{8}$$

where $\mathbf{L}_{\mathbf{A}}\mathbf{f}(x)$, $\mathbf{L}_{\mathbf{C}}^{\mathbf{y}}\mathbf{f}(\mathbf{y})$, $\mathbf{d}^{\mathbf{T}}\mathbf{f}(x)$ can be regarded as the similarity components for x, while $\mathbf{L}_{\mathbf{B}}\mathbf{f}(\mathbf{y})$, $\mathbf{L}_{\mathbf{C}}^{\mathbf{y}}\mathbf{f}(\mathbf{y})$, $\mathbf{e}^{\mathbf{T}}\mathbf{f}(\mathbf{y})$ accordingly for $\mathbf{y}$. These similarity components are models as the weights that connect neurons of the last two layers in Fig. 2.

In Fig. 2, $\mathbf{f}(x)$ and $\mathbf{f}(\mathbf{y})$ are fed to two branches of similarity sub-network, and each bran includes a fully-connected layer. We divide the activations of these two layers into six parts according to the six similarity components of Eq. (8). In the top branch the neural layer connects to $\mathbf{f}(x)$ and outputs $\mathbf{L}_{\mathbf{A}}\mathbf{f}(x)$, $\mathbf{L}_{\mathbf{C}}^{x}\mathbf{f}(x)$, $\mathbf{d}^{\mathbf{T}}\mathbf{f}(x)$, respectively. In the bottom branch, the layer outputs $\mathbf{L}_{\mathbf{B}}\mathbf{f}(\mathbf{y})$, $\mathbf{L}_{\mathbf{C}}^{\mathbf{y}}\mathbf{f}(\mathbf{y})$, $\mathbf{e}^{\mathbf{T}}\mathbf{f}(\mathbf{y})$, respectively, by connecting to $\mathbf{f}(\mathbf{y})$. In this way, the similarity measure is tightly integrated with the feature representations, and they can be jointly optimized during the model training.

The objective of this deep similarity learning is to seek a function $S(\mathbf{f}(x), \mathbf{f}(\mathbf{y}))$ that satisfies a set of similarity constraints. Recall that $(\mathbf{f}(x), \mathbf{f}(\mathbf{y}))$ are the feature representations for samples of image and video, and we use $\mathbf{W}$ to indicate their parameters, and denote $\Phi = (\mathbf{L}_{\mathbf{A}}, \mathbf{L}_{\mathbf{B}}, \mathbf{L}_{\mathbf{C}}^{\mathbf{x}}, \mathbf{L}_{\mathbf{C}}^{\mathbf{y}}, \mathbf{d}, \mathbf{e}, f)$ as the similarity components for sample matching. Assume that $\{(\{x_i, \mathbf{y}_i\}, \ell_i)\}_{i=1}^N$ is a training set of image-to-video sample pairs, and $\ell_i = 1$ denotes the corresponding label of $\{x_i, \mathbf{y}_i\}$ indicating x_i and $\mathbf{y}_i$ are from the same class, $\ell_i = -1$ otherwise.

The deep similarity model is expected to satisfy the following hinge-like loss function:

$$(\mathbf{W}, \Phi) = \arg\min_{\mathbf{W},\Phi} \sum_{i=1}^{N} (1 - \ell_i S(\mathbf{f}(x_i), \mathbf{f}(\mathbf{y}_i))_+ + \Psi(\mathbf{W}, \Phi) \tag{9}$$

where $\Psi(\mathbf{W}, \Phi) = \lambda\|\mathbf{W}\|^2 + \mu\|\Phi\|^2$ denotes the regularizer on the parameters of the feature representation and generalized similarity models in Eq.(9).

The proposed framework can be easily trained with standard back propagation method. We establish an end-to-end framework for image-to-video person re-id by performing the feature extraction and similarity learning simultaneously. By incorporating LSTM connections between the CNN and similarity network, we aim to improve the performance of image-to-video person re-id systems using temporal information present in the video sequence.

D. Implementation Details

1) CNN Models: In the proposed framework, we use the CNN model to extract the features of an input image and the features of each frame of video. We use the similar architecture as AlexNet [25] without the fully-connect layers. The vectorized representation of the final layer activation map is output as the features of an image. For a single image, the CNN will output a 400 dimension feature vector $\mathbf{f}(x)$. The same architecture of CNN is used for the feature extraction of each frame of a video sequence. For each video frame, CNN will output a 400 dimension feature vector, which is passed forward to the LSTM layers to further encode the spatial information of video.

2) LSTM Network: We use the LSTM network to further encode the temporal information of video sequence. For LSTM, we follow the architecture of [51]. The input of LSTM is the output of CNN. We adopt two LSTM layers for the video information embedding. One layer is an encoder layer, and the other is decoder layer. Suppose the length of a video is T, the LSTM network will output a $400 \times T$ dimension feature map. We then apply a $1 \times T$ convolutional kernel on the feature map, and finally get a 400 dimension feature vector $\mathbf{f}(\mathbf{y})$ of video according to Eq.(5). In this paper, T is set to 10 for convenience.

3) Similarity Sub-Network: As shown in Fig. 2, we use a 400×801 fully connected layer to construct the similarity sub-network. The component related to input feature $\mathbf{f}(x)$, i.e., $\mathbf{L}_{\mathbf{A}}\mathbf{f}(x)$, $\mathbf{L}_{\mathbf{C}}^{x}\mathbf{f}(x)$, $\mathbf{d}^{\mathbf{T}}\mathbf{f}(x)$, are with the dimension of 400, 400, and 1, respectively. The component related to

TABLE I
THE SUMMARY OF DATASETS OF PRID-2011, iLIDS-VID, AND MARS

Dataset	PRID-2011	iLIDS-VID	MARS
#ID	200	300	1,261
#Tracklects	400	600	20,478
#Bboxes	40,000	43,800	1,191,003
#Cameras	2	2	6
Label	DPM+GMMCP	hand	hand

input feature $\mathbf{f}(\mathbf{y})$, i.e., $\mathbf{L_B f}(\mathbf{y})$, $\mathbf{L_C^y f}(\mathbf{y})$, $\mathbf{e^T f}(\mathbf{y})$, are with the dimension of 400, 400, and 1, respectively.

4) Data Augmentation: To increase the diversity of the training sequences, data augmentation including cropping and mirroring is applied. For a given sequence, the same augmentation is applied to all frames. During the testing phase, data augmentation is also applied, and the similarity scores between sequences are averaged over all the augmentation conditions.

IV. EXPERIMENTAL RESULT

In this section, we present the evaluation results of the proposed approach. We first describe the data setup in this paper. Comparison with several state-of-the-art person re-id methods is also presented. In addition, the ablation study is described to provide more insights of the proposed method.

A. Dataset Setup

In this paper, we adopt three datasets to evaluate our proposed image-to-video person re-id method as shown in Fig. 3, i.e., PRID-2011 [53], iLIDS-VID [54], and MARS [45]. A brief summary of three dataset are list in Table I.

PRID-2011 dataset includes 400 image sequences for 200 persons from two adjacent camera views. Each sequence is between 5 and 675 frames, with an average of 100. Compared with iLIDS-VID, this dataset less challenging due to being captured in uncrowded outdoor scenes with rare occlusions and clean background. However, the dataset has obvious color changes and shadows in one of the views. Similar to the protocol used in [48], we only use the first 200 persons appearing in both cameras for evaluation.

iLIDS-VID dataset consists of 600 image sequences for 300 randomly sampled people, which was created based on two non-overlapping camera views from the i-LIDS multiple camera tracking scenarios. The sequences are of varying length, ranging from 23 to 192 images, with an average of 73. This dataset is very challenging due to variations in lighting and viewpoint caused by cross-camera views, similar appearances among people, and cluttered backgrounds.

MARS dataset is a newly released dataset for video-based person re-id. It is an extension of the Market-1501 dataset [18], which is captured by six near-synchronized cameras in the campus. There were Five $1,080 \times 1,920$ HD cameras and one 640×480 SD camera. MARS consists of 1,261 different pedestrians whom are captured by at least 2 cameras. There are 20,478 tracklets, with 1,191,003 bounding box produced by DPM [36] and GMMCP [55] descriptors.

B. Experimental Setting and Evaluation Protocol

Following the setting of [48], for the experiment on the dataset of PRID-2011 and iLIDS-VID, the data was randomly split into 50% of person for training and 50% of person for the test. All experiments were repeated 10 times with different test/train splits and the results averaged to ensure a stable result. For MARS, we used the presetting of training/test split. The dataset used in this paper were all video dataset. As in our work, we focus on the image-to-video person re-id, in which the probe is an image. During the training for PRID-2011 and iLIDS-VID, we randomly selected video and used its first frame as the input image, and video of the other viewpoint as the input video sequence. For MARS, as there were many viewpoint videos for a single person, we randomly selected the video of one viewpoint and used its first frame as the input image, and randomly selected a video of another viewpoint as the input video of our network. The same strategy was adopted during the test. Although LSTM can handle arbitrary length video sequence, we select 10 consecutive frames of video for convenience. The margin of our network was set to 1 and trained for 500 epochs using stochastic gradient descent with the learning rate of 1e-4. An epoch consists of showing all positive image-to-video pairs and an equal number of negative pairs random sampled from all training persons. All experiments were conducted on a desktop computer with a NVIDIA Titan X GPU.

In experiments, for each pedestrian, the matching of his or her probe image (captured by one camera) with the gallery sequence (captured by other cameras) is ranked. To reflect the statistics of the ranks of true matches, the average Cumulative Match Characteristic (CMC) over 10 trails was adopted as the evaluation metric. Specifically, in the testing phase, the similarity between probe image features and those of gallery sequences were computed firstly. Then, for each probe person, a rank order of all the persons in the gallery was sorted from the one with the smallest distance to the biggest distance. Finally, the percentage of true matches found among the first m ranked persons was computed and denoted as rank(m).

C. Experimental Results

1) Results of Proposed Method: With the above setting, we evaluated the proposed framework for image-to-video person re-id. The CMC was adopt as the evaluation metric. Table II shows the experimental results of our work on three datasets, i.e., PRID-2011 [53], iLIDS-VID [54], and MARS [45], respectively. On PRID-2011 [53], our method achieved a rank1 accuracy of 68.5%, and the results on iLIDS-VID [54] and MARS [45] were 39.5% and 56.5%, respectively.

In order to provide more insights on the performance of our approach, we conduct ablation studies by isolating the LSTM layer and similarity sub-network as two variations of our methods and evaluated their performance on three datasets, respectively. In the first variant, we removed the LSTM network in the proposed framework and directly used the extracted CNN features of each frame of video sequences. Average pooling was performed on the features of each video

TABLE II

RESULTS OF THE ABLATION STUDIES. ACCURACIES ARE PRESENTED BY CMC IN RANK 1, 5, 20, RESPECTIVELY

Methods	PID-2011			iLIDS-VID			MARS		
Rank r	1	5	20	1	5	20	1	5	20
Baseline1	64.7%	82.3%	93.5%	36.1%	61.6%	83.1%	46.8%	66.7%	76.9%
Baseline2	66.5%	83.7%	94.6%	36.5%	64.7%	83.4%	51.4%	67.3%	77.4%
Baseline3	65.3%	82.5%	93.8%	35.5%	62.3%	83.1%	49.4%	66.9%	77.2%
Ours(full)	**68.5%**	**84.7%**	**96.3%**	**39.5%**	**66.9%**	**86.5%**	**56.5%**	**70.6%**	**79.8%**

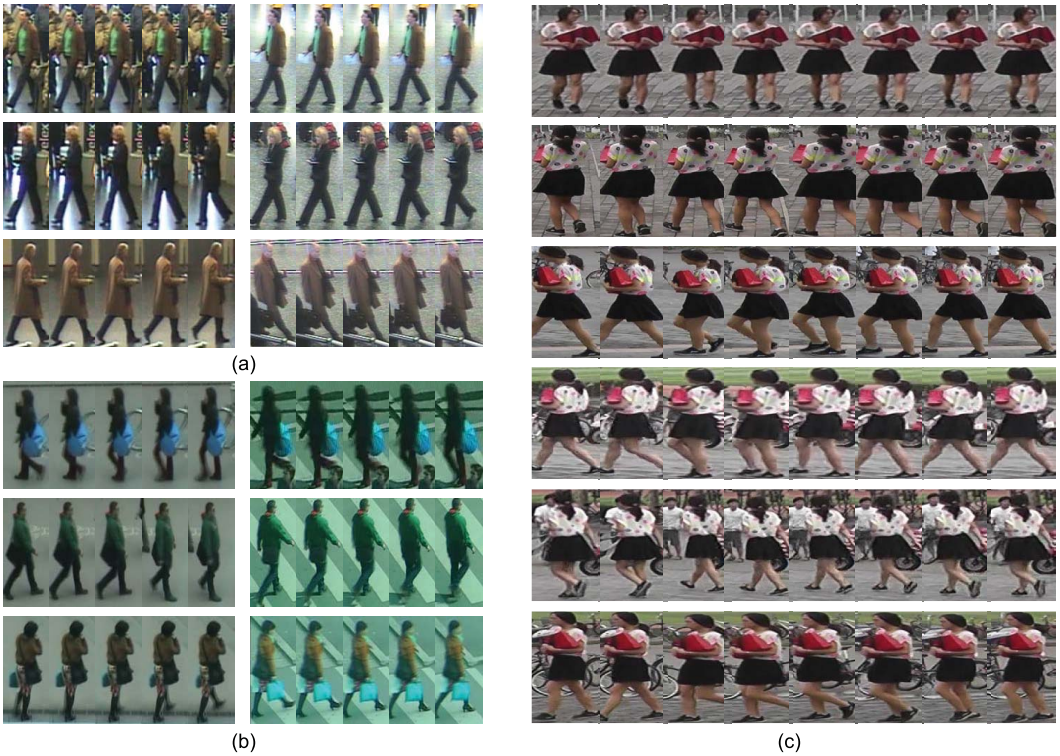

Fig. 3. Examples of the video sequence of three datasets used in this paper. (a) and (b) are example pairs of the image sequence of the same pedestrian in different camera views from iLIDS-VID, and PRID-2011, while (c) shows the six video sequences of the same person from MARS dataset.

frame to output a feature vector which has the same length with the feature vector of probe image. We then invoked the similarity sub-network for the distance metric learning. By using this strategy, we wanted to evaluate the effectiveness of LSTM network in the proposed framework. This method was remarked as Baseline1. In the second variant method, we removed the similarity sub-network. We used the CNN and LSTM network to extract the features of probe image and video sequence. We then used the XQDA [17] metric for the similarity evaluation. By using this strategy, we wanted to evaluate the effectiveness of the similarity sub-network in our framework. The variant was remarked as Baseline2. Besides, we also used the Euclidean distance to measure the similarity between the feature of probe image and video, which was remarked as Baseline3.

The comparison results of different variants are listed in Table II. From the table, we can see that the full version of our proposed framework, Ours(full), achieved the highest accuracies on all the three datasets.

In Table II, Compared with Baseline1, the improvement of $rank1$ accuracy of Ours(full) were 3.8%, 3.4%, 5.7% on the datasets of PID-2011, iLIDS-VID and MARS respectively. As compared with Ours(full), Baseline1 was the variant without LSTM network, the results showed that the LSMT network was effective in further encoding the information of video in the proposed framework. Compared with the variants of Baseline2 and Baseline3, Ours(full) has also much improved on all three datasets, which implied that the embedded similarity sub-network in our framework was also effective in learning the distance metric. From the comparison results of Table II, we can see that the full version of our method was the best on all the three datasets, which showed the effectiveness of the joint feature representation and distance metric learning in our framework.

2) Comparison With the State-of-the-Arts: We also compared the performance of our proposed approach with several state-of-art methods from the literature by using CMC for evaluation, which was averaged over ten training/test

TABLE III

RESULTS OF THE STATE-OF-THE-ART METHODS ON THE PID-2011 AND iLIDS-VID DATASETS. ACCURACIES ARE PRESENTED BY CMC IN RANK 1, 5, 10, AND 20. (FOR CNN+XQDA [46], RCN [48], RFA-NET [49] THE RESULTS REPORTED IN THEIR PAPERS WERE BASED ON THE VIDEO-TO-VIDEO PERSON RE-ID. THE RESULTS REPORTED HERE WERE OBTAINED BY USING THEIR MODEL WITH OUR IMAGE-TO-VIDEO SETTING.)

Methods	**PID-2011**				**iLIDS-VID**			
Rank R	1	5	10	20	1	5	10	20
Color+DVR [53]	41.8%	63.8%	84.5%	88.3%	32.7%	56.5%	73.2%	77.4%
SDALF [56]+DVR [53]	31.6%	58%	81.3%	85.3%	26.7%	49.3%	65.4%	71.6%
BoW+XQDA [17]	31.8%	58.5%	78.2%	81.9%	14.0%	32.2%	52.1%	59.5%
Hog3D [57]+XQDA [17]	21.7%	51.7%	83.5%	87.0%	16.1%	41.6%	69.2%	74.5%
CNN+XQDA [46]	66.3%	**85.1%**	92.5%	96.5%	37.8%	63.7%	77.6%	84.9%
RCN [48]	54.3%	73.4%	88.5%	92.5%	28.8%	57.7%	69.6%	81.9%
RFA-Net [49]	68.3%	81.1%	92.7%	96.8%	**39.6%**	65.8%	78.3%	85.3%
Ours(full)	**68.5%**	84.7%	**93.1%**	**97.3%**	39.5%	**66.9%**	**79.6%**	**86.6%**

TABLE IV

RESULTS OF THE STATE-OF-THE-ART METHODS ON THE DATASET OF MARS. ACCURACIES ARE PRESENTED BY PRECISION IN RANK 1, 5, AND 20. (FOR CNN+XQDA [46], RCN [48], RFA-NET [49] THE RESULTS REPORTED IN THEIR PAPERS WERE BASED ON THE VIDEO-TO-VIDEO PERSON RE-ID. THE RESULTS REPORTED HERE WERE OBTAINED BY USING THEIR MODEL WITH OUR IMAGE-TO-VIDEO SETTING.)

Methods	**MARS**		
Rank R	1	5	20
HistLBP+XQDA [17]	18.6%	33.0%	45.9%
LOMO+XQDA [17]	30.7%	46.6%	60.9%
BoW+KISSME [23]	30.6%	46.2%	59.2%
SDALF+DVR [53]	4.1%	12.3%	25.1%
HOG3D [57]+KISSME [23]	2.6%	6.4%	12.4%
GEI [58]+KISSME [23]	1.2%	2.8%	7.4%
CNN+XQDA [46]	50.3%	67.5%	80.1%
RCN [48]	44.2%	58.6%	77.6%
RFA-Net [49]	55.6%	69.3%	82.8%
Ours(full)	**56.5%**	**70.6%**	**83.5%**

partitions. To ensure a fair comparison, all the methods were trained and tested using the same datasets and same training/test split. We used the same partition rule as in [53] for the training/test sets.

In Table III, we compared our approach with several state-of-the-art methods on the datasets of PRID-2011 and iLIDS-VID. Four descriptors were compared, i.e., color, color+LBP, SDALF [56], Salience [59], and BoW [18]. Three metric learning methods, i.e., DVR [53], XQDA [17], and KISSME [23] were evaluated. We also list the results on the two dataset of two video-based person re-id methods, including RCN [48][1] and RFA-Net [49].[2] As the original models of the two methods were based on the matching of video-to-video, we modified their models according to our setting for image-to-video person re-id. Thus, the reported results in Table III and Table IV were not the same as those in their original papers.

[1]https://github.com/niallmcl/Recurrent-Convolutional-Video-ReID

[2]https://github.com/daodaofr/caffe-re-id

From Table III, we can see that our proposed method achieved the rank1 accuracy of 68.5% and 39.5% on the PID-2011 and iLIDS-VID, respectively. The results of RFA-Net [49] were comparable to our approach. On the MARS dataset, results of another set of features were presented, i.e., HistLBP [60], LOMO [17], BoW [18], and SDALF [56]. Table IV showed the experimental results. The $rank1$ accuracy of proposed method on MARS was 56.5%. The comparison results showed that our proposed approach was comparable to the state-of-the-arts and was effective for the image-to-video problem.

V. CONCLUSION

Image-to-video person re-id is different from the most existing research of person re-id in probe and gallery setting, where the probe is an image of a person and the gallery is consists of videos of many persons with different viewpoints. In this paper, we propose a novel framework for image-to-video person re-id problem, in which the feature representation and distance metric learning are jointly performed and optimized. We adopt CNN to extract the feature of probe image. For the video, CNN models and LSTMs are used to extract the features of person's appearance and embed temporal information in the video sequence. Finally, a similarity sub-network is adopted to learn the distance metric to measure the similarity of the input image and gallery videos. Experimental results on three challenging person re-id datasets demonstrate that the proposed method achieves the state-of-the-art.

REFERENCES

[1] T. D'Orazio and G. Cicirelli, "People re-identification and tracking from multiple cameras: A review," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1601–1604.

[2] B. Wang, G. Wang, K. L. Chan, and L. Wang, "Tracklet association with online target-specific metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1234–1241.

[3] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, Jan. 2013.

[4] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[5] W. Zhang, B. Ma, K. Liu, and R. Huang, "Video-based pedestrian re-identification by adaptive spatio-temporal appearance model," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2042–2054, Apr. 2017.

[6] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1345–1353.

[7] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jun. 2006, pp. 1528–1535.

[8] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 262–275.

[9] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2666–2672.

[10] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3610–3617.

[11] X. Zhou, N. Cui, Z. Li, F. Liang, and T. S. Huang, "Hierarchical gaussianization for image classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1971–1977.

[12] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3318–3325.

[13] W. Zuo, D. Ren, D. Zhang, S. Gu, and L. Zhang, "Learning iteration-wise generalized Shrinkage–Thresholding operators for blind deconvolution," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1751–1764, Apr. 2016.

[14] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3550–3557.

[15] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 536–551.

[16] L. Lin, X. Wang, W. Yang, and J.-H. Lai, "Discriminatively trained And-Or graph models for object shape detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 959–972, May 2015.

[17] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2197–2206.

[18] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1116–1124.

[19] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 209–216.

[20] X.-Y. Jing *et al.*, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 695–704.

[21] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1335–1344.

[22] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2009, pp. 498–505.

[23] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2288–2295.

[24] S. Liao and S. Z. Li, "Efficient PSD constrained asymmetric metric learning for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3685–3693.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[27] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[28] X. Liang *et al.*, "Human parsing with contextualized convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1386–1394.

[29] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, "A deep structured model with Radius–Margin bound for 3D human activity recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 256–273, Jun. 2016.

[30] L. Lin, Y. Lu, C. Li, H. Cheng, and W. Zuo, "Detection-free multiobject tracking by reconfigurable inference with bundle representations," *IEEE Trans. Cybern.*, vol. 46, no. 11, pp. 2447–2458, Nov. 2016.

[31] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1288–1296.

[32] M. Li, Q. Wang, D. Zhang, P. Li, and W. Zuo, "Joint distance and similarity measure learning based on triplet-based constraints," *Inf. Sci.*, vols. 406–407, pp. 119–132, Sep. 2017.

[33] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.

[34] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3908–3916.

[35] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

[36] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[37] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 34–39.

[38] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[39] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, Oct. 2015.

[40] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[41] R. Satta. (2013). "Appearance descriptors for person re-identification: A comprehensive review." [Online]. Available: https://arxiv.org/abs/1307.5748

[42] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 475–491.

[43] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 791–808.

[44] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[45] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler, "Re-identification of pedestrians in crowds using dynamic time warping," in *Proc. Comput. Vision ECCV Workshops Demonstrations*, 2012, pp. 423–432.

[46] L. Zheng *et al.*, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.

[47] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen. (2015). "Temporal pyramid pooling based convolutional neural networks for action recognition." [Online]. Available: https://arxiv.org/abs/1503.01224

[48] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1325–1334.

[49] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 701–716.

[50] J. Wang and A. L. Yuille, "Semantic part segmentation using compositional model combining shape and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1788–1797.

[51] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 6645–6649.

[52] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1089–1102, Jun. 2017.

[53] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 688–703.
[54] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. Scandin. Conf. Image Anal.*, 2011, pp. 91–102.
[55] A. Dehghan, S. Modiri Assari, and M. Shah, "GMMCP tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4091–4099.
[56] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2360–2367.
[57] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. 19th Brit. Mach. Vis. Conf.*, 2008, pp. 275–281.
[58] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.
[59] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.
[60] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.

Dongyu Zhang received the B.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2003 and 2010, respectively. He is currently a Research Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. His current research interests include computer vision, machine learning, deep learning, and related multimedia applications.

Wenxi Wu is currently pursuing the master's degree in computer vision and deep learning with the School of Data and Computer Science, Sun Yat-Sen University. He is the Kaggle Gold Metal Owner of the Nature Conservancy Fisheries Monitoring Competition and received the 11th-place in ILSVRC2015 (DET).

Hui Cheng received the B.Eng. degree in electrical engineering from Yanshan University, Qinhuangdao, China, the M.Phil. degree in electrical and electronic engineering from the Hong Kong University of Science and Technology, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong. She was a Post-Doctoral Fellow with The Chinese University of Hong Kong from 2006 to 2007. She is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou. Her current research interests include intelligent robots and networked control systems.

Ruimao Zhang received the B.E. and Ph.D. degrees from Sun Yat-sen University, Guangzhou, China, in 2011 and 2016, respectively. From 2013 to 2014, he was a visiting Ph.D. student with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. He is currently a Post-Doctoral Fellow with the Department of Electronic Engineering, The Chinese University of Hong Kong. His current research interests include computer vision, deep learning, and related multimedia applications. He currently serves as a Reviewer of several academic journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *Pattern Recognition*, and *Neurocomputing*.

Zhenjiang Dong received the M.S. degree in telecommunication from the Harbin Institute of Technology in 1996. He is the Deputy Head of the Service Institute of ZTE Corporation, the Standing Director of Chinese Association for Artificial Intelligence. His current research interests include cloud computing, mobile Internet, natural language processing, and multimedia analysis.

Zhaoquan Cai received the Ph.D. degree in computer science from the Huazhong University of Science and Technology. He is currently a Professor of Computer Science with Huizhou University, China. He His current research interests include computer networks, intelligent computing, and database systems.