Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09398-0

T-MAN: a neural ensemble approach for person re-identification using spatio-temporal information

Nirbhay Kumar Tagore[1] · **Pratik Chattopadhyay**[1] · **Lipo Wang**[2]

Received: 17 January 2020 / Revised: 10 July 2020 / Accepted: 21 July 2020 /
Published online: 03 August 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Person re-identification plays a central role in tracking and monitoring crowd movement in public places, and hence it serves as an important means for providing public security in video surveillance application sites. The problem of person re-identification has received significant attention in the past few years, and with the introduction of deep learning, several interesting approaches have been developed. In this paper, we propose an ensemble model called Temporal Motion Aware Network (T-MAN) for handling the visual context and spatio-temporal information jointly from the input video sequences. Our methodology makes use of the long-range motion context with recurrent information for establishing correspondences among multiple cameras. The proposed T-MAN approach first extracts explicit frame-level feature descriptors from a given video sequence by using three different sub-networks (*FPAN*, *MPN*, and *LSTM*), and then aggregates these models using an ensemble technique to perform re-identification. The method has been evaluated on three publicly available data sets, namely, the *PRID-2011*, *iLIDS-VID*, and *MARS*, and re-identification accuracy of 83.0%, 73.5%, and 83.3% have been obtained from these three data sets, respectively. Experimental results emphasize the effectiveness of our approach and its superiority over the state-of-the-art techniques for video-based person re-identification.

Keywords Spatio-temporal information · Ensemble model · Person re-identification · Deep learning

✉ Pratik Chattopadhyay
pratik.cse@iitbhu.ac.in

Nirbhay Kumar Tagore
nirbhaykrtag.rs.cse17@iitbhu.ac.in

Lipo Wang
ELPWang@ntu.edu.sg

[1] Pattern Recognition Lab, Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi, 221005, Uttar Pradesh, India

[2] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, Singapore

1 Introduction

Person re-identification is a process of establishing one-one correspondence among the images of individuals captured by non-overlapping cameras at different points of time. A basic re-identification system can be broadly segregated into three phases, i.e., person detection, tracking, and final retrieval. The problem of person re-identification has got expanding consideration [58] [27], which intends to identify an individual captured by one camera in the field of view of another camera positioned at a different place. Computer vision-based person re-identification must be robust against variation of lighting, postures, perspectives, etc. Also, the continuous recording of videos from camera network results in a huge volume of data, manual monitoring of which is time-intensive and error-prone. Hence, there is a high demand for developing automated person re-identification algorithms that can be readily deployed in practical sites. Our work focuses on developing a re-identification algorithm that can meet these aforementioned challenges effectively.

Here, we propose to design a framework that can jointly handle the texture features along with pose and spatio-temporal information from the video sequences of individuals to perform person re-identification. The main contributions of the paper are as follows:

1. Introducing novel motion-based features and proposing a new ensemble architecture termed as Temporal Motion Aware Network (*T-MAN*) for person re-identification.
2. Carrying out temporal modelling of motion information by employing an ensemble of deep neural network models, namely, (i) Full-Body Pose Attention Network (*FPAN*), (ii) Motion Pooling Network (*MPN*), and (iii) Long-Short Term Memory Neural Network (*LSTM*).
3. Performing extensive evaluation and comparative analysis with state-of-the-art approaches using three public benchmark data sets to establish the effectiveness of our approach.

The structure of the rest paper is as follows. In Section 2, we discuss previous approaches related to person re-identification with a major focus on deep learning-based models. In the following section, i.e., Section 3, we explain the proposed approach along with the network architecture in detail. Section 4 focuses on the experimental results and analysis, as well as comparison with competing approaches. Finally, we conclude our paper in Section 6 and point out future research scopes.

2 Related work

Person re-identification approaches from still images primarily focus on metric learning and feature representation techniques that are invariant to change of viewpoint and other physical conditions like illumination and occlusion. Existing research in this domain can be broadly classified into the following categories: *image-based person re-identification* and *video-based person re-identification*, which are discussed in the next two sub-sections.

2.1 Image-based person re-identification

In [15], a view-invariant re-identification approach has been described that takes into consideration the spatial information along with color features in image frames by ensembling classifier predictions with discriminant localized features. Since maximum spatial information about a subject can be obtained from silhouette regions closer to the body axes of

symmetry, the authors in [10] computed symmetry and asymmetry perceptual attributes of visual features to effectively perform re-identification. The work in [5] focuses on extracting color-based features from body parts, i.e., chest, head, thighs, and legs, to handle the pose variation problem in person re-identification. Kviatkovsky et al. [21] proposed an illumination invariant feature representation technique using log-chromaticity attributes. This approach has been shown to perform satisfactorily in the presence of varying lighting conditions. The approach given in [31] creates a global representation of images by transforming the neighborhood descriptors into the Fisher discriminant space. In [43], distance metric learning has been employed after feature extraction for emphasizing the inter-person distance while simultaneously de-emphasizing the intra-person distance. Here, a large margin nearest neighbor metric (LMNN) has been used as an improvement over the traditional *k-Nearest Neighbor* classification technique. In [37], Prosser et al. use an ensemble of Rank-SVMs to learn pairwise similarity and, henceforth, formulate a ranking problem. In [54], a Relative Distance Comparison (RDC) scheme has been proposed based on a soft discriminative scheme for large and small distances corresponding to incorrect matches and correct matches, respectively. An unsupervised Bag-of-Words descriptor for person re-identification as well as a new benchmark data set termed as *Market-1501* has been introduced in [55]. However, this data set contains only a few image frames from each individual, but not a complete video sequence. Since we aim to construct motion-based features for person re-identification, this data set has not been used in the study. In [45], a comparative study of several classification methods for person re-identification task has been presented, namely, the regularized Pairwise Constrained Component Analysis, Kernel Local Fisher Discriminant Analysis, Marginal Fisher Analysis, and a ranking ensemble voting scheme. A new feature termed as the *Local Maximal Occurrence* (LOMO) is derived and a *Cross View Quadratic Analysis* metric is learned for person re-identification in [24]. The authors of [20] present a simple yet effective distance metric learning strategy from equivalence constraints which can work even in the absence of labelled data. The introduction of deep learning has significantly benefited research on Computer Vision-based person tracking including detection, recognition, re-identification [4, 30, 36, 39, 48]. The use of Deep learning in person re-identification started with the work of [22], following which several improved deep learning approaches to solve the same problem have also been developed. Among this, a Deep Siamese network-based approach has been proposed in [2] in which two parallel convolution networks are used that are tied with weights to generate the feature descriptors at its two channels. These features are next compared to predict if the given pair of images are the same or different.

2.2 Video-based person re-identification

Previous work on temporal modelling methods on video-based person re-identification use either Recurrent Neural Network (RNN) models, or temporal attention-based models. In [33], McLaughlin et al. first introduced the concept of modelling temporal information between frames by Recurrent Neural Network (RNN), in which the average of RNN cell outputs have been used as clip level representations. Like [33], Yan et al. [47] also employed RNN to encode sequence features and considered the last hidden state to preserve the entire video information. Liu et al. [27] presented a Quality Aware Network (QAN), which is an attention weighted average to compute temporal features, where the attention scores are created from frame-level feature maps. The approaches described in [57] and [46] extract attention features as well as temporal RNN-based features to preserve the dynamic motion information. The two-stream network developed by Chung et al. in [6] computes features

from both RGB images as well as optical flow, and uses simple temporal pooling to aggregate the features descriptors. In [34], a classification ensembling approach is discussed by fusing a number of deep networks to improve the generalization. Another end-to-end trainable architecture, namely, the Accumulative Motion Context (AMOC) has been proposed in [25] to jointly handle the appearance representation and motion context present in a given video sequence. In [49], an unsupervised approach for label estimation is presented based on a dynamic graph matching (DGM) framework to improve the label estimation process in person re-identification. Here, intermediate labels have been used to iteratively refine the graph structure for labelling the data. The work in [23] describes an unsupervised approach that can jointly learn from camera tracklet association and cross-camera tracklet correlation to improve the scalability of the model without the requirement of rigorous manual labelling. In [40], re-identification is performed by identifying the most discriminative features in an image sequence, whereas in [26], pose-based alignment of image frames in a video is carried out using spatio-temporal appearance features before the subsequent person re-identification step. Re-identification from video sequences is done in [19] by computing local descriptors based on 3D spatio-temporal gradients. The *Temporally Aligned Pooling Feature Representation* (TAPR) computed in [14] extracts motion information from the video sequences by tracking super-pixels at the lowest portions of human beings. In [18], the re-identification problem is formulated as a block sparse recovery problem which is then solved using alternating directions framework. The work in [52] introduces top-push distance learning model for video-based person re-identification to overcome the challenges due to change of pose and camera view-point, occlusion, and lighting variation. Another covariance descriptor for face verification and person re-identification has been described in [32] that can handle both background and illumination variations. The work in [56] employs a *Confidence Weighted Similarity* (CWS) for similarity measurement and a cascaded fine-tuning strategy to carry out the classification process.

2.3 Attention-based models

Since, in this work we derive attention-based features from the input video sequences, we next review some recently developed attention-based methods that have been used in computer vision tasks such as image segmentation, recognition, etc. Recent use of attention features can be found in fine-grained image recognition methods such as [13, 53]. The work in [13] presents a mutually reinforced way to jointly learn the discriminative region attention and region-based feature representation, while that in [53] describes a similar approach by proposing a multi-attention convolutional neural network consisting of three sub-modules, namely convolution, channel grouping, and part-classification to handle the localization and part-based fine-grained feature learning. Attention-based models have also been used in [9, 29] for video object detection and segmentation. The authors of [29] present a new Siamese model termed as *Co-attention Siamese Network* (COSNet) to process multiple reference frames together for segmentation and encoding of useful image features. On the other hand, [9] addresses the challenge of saliency shift by introducing a new model equipped with saliency-shift-aware *conv-LSTM*. An adaptive region proposal scheme for object detection is given in [28] that uses trainable correlation filters to develop a two-stream framework to distinguish between background and foreground targets and help in accurate segmentation.

From the extensive literature survey, it is seen that there does not exist suitable techniques that jointly handle both the spatio-temporal as well as attention-based features to solve the problem of person re-identification. In this paper, we propose a new architecture based on

a stacked ensemble model of three deep networks to improve over other state-of-the-art re-identification techniques.

3 Proposed work

In this work, we follow an ensemble-based classification scheme to re-identify individuals in the fields of view of two cameras. Specifically, predictions from three different deep neural networks are fused to estimate the class of a test subject. The three sub-networks used in the ensemble model are (i) a *Full-Body Pose Attention Network* (*FPAN*), (ii) a *Motion Pooling Network* (*MPN*), and (iii) a Convolutional *Long-Short Term Memory Network* (*LSTM*). These three models are based on the popular *ResNet-101* architecture [16], and preserve information related to the different aspects of human motion. While the *FPAN* captures mainly appearance-related information of an individual, the *MPN* captures dominant motion features, and the *LSTM* derives dynamic information from the spatial correlation between frames in a captured sequence. Use of *ResNet-101* as the base network is justified since its effectiveness in object detection and recognition has already been well-established [7, 44]. The pre-trained version of the *ResNet-101* architecture [38] has been used here to generate the frame-level feature descriptors from the input image sequences. The complete re-identification approach and detailed discussion on the above-mentioned sub-networks are given next. The overall pipeline of the proposed re-identification approach consists mainly of three modules: (i) training of the individual sub-networks, namely, *FPAN*, *MPN*, and *LSTM*, (ii) aggregation for features from the trained models, and (iii) predicting the class of a test subject, and is explained with the help of a signal-flow diagram, as shown in Fig. 1.

With reference to the figure, initially, the entire video is segmented into non-overlapping clips of T frames. Each set of T frames present in a clip is next passed one-by-one through a pre-trained *ResNet-101* model, and these T *ResNet* features are input to each of the three sub-networks, i.e., *FPAN*, *MPN*, and *LSTM*, to compute clip-level motion features. Similarly, clip-level features are computed from each of the other clips at the three sub-networks, and finally, these clip-level features at each sub-network are aggregated to obtain three different features from an input video sequence, which preserve important kinematic characteristics of human motion. Figure 1 shows separate *ResNet-101* blocks to make it easier for the readers to understand the flow of the work. During implementation, a single such *ResNet* block may be considered, and each frame may be passed separately through that network to obtain the deep features corresponding to that frame. Let the dimensions of each feature map at the final convolution layer of the *ResNet-101* model be $w \times h$. Since, this layer contains 2048 feature maps, the tensor size at this layer can thus be represented by $[w, h, 2048]$. Let us denote the tensor corresponding to the t^{th} frame of clip c by f_c^t, where t = 1,2,..., T and c= 1,2,...$\mathcal{C}$. If there are $\mathcal{C}$ total clips, each of *FPAN*, *MPN*, and *LSTM* fuses information from the $\mathcal{C}$ clips to compute class probabilities F_1, F_2, and F_3, which are finally fused to predict the final class.

Full-Body Pose Attention Network In Full-Body Pose Attention Network (*FPAN*), the average of attention scores corresponding to different clips (i.e., fragments of an input sequence) are computed. This is done by employing a temporal attention layer after the final convolution layer of the *ResNet-101* as explained next. To compute the attention score for

clip c, we consider the attention tensor b_c^t as the average of the T ResNet-101 generated tensors $\{f_c^1, f_c^2, ..., f_c^T\}$ as shown in (1):

$$b_c^t = \frac{1}{T}\sum_{t=1}^{T} f_c^t. \tag{1}$$

The size of tensor b_c^t can also be represented as $[w, h, 2048]$. Next, we multiply each of the T *ResNet-101* tensors, $f_c^1, f_c^2, ..., f_c^T$, with the attention tensor b_c^t, element-wise, and sum up all the resultant tensors to obtain a single attention-infused tensor for clip c denoted by m_c. Mathematically,

$$m_c = \sum_{t=1}^{T} b_c^t \otimes f_c^t, \tag{2}$$

where $\otimes$ denotes the tensor product operator. The size of tensor m_c is also $[w, h, 2048]$. The feature maps obtained after the multiplication operation are now passed through a convolution layer with 256 kernels, each of dimensions $w \times h$ to reduce the tensor size to $[w, h, 256]$, following which a fully connected layer with a single node is considered that computes the attention score for clip c. If this score is denoted by S_c, then the final attention score (S)

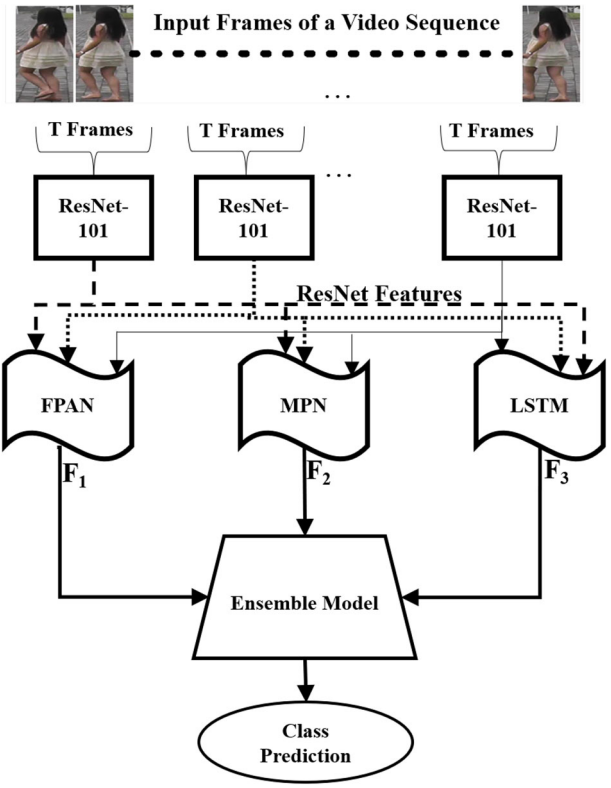

Fig. 1 A signal flow diagram of the proposed approach

provided by the *FPAN* sub-network is obtained by summing up the attention scores from each of the $\mathcal{C}$ clips as shown in (3):

$$S = \sum_{c=1}^{\mathcal{C}} S_c. \tag{3}$$

Now, if there are N subjects in the gallery set, then we compute N such scores following (3). Let us denote these N scores by S^1, S^2, ..., S^N. If the attention score of an input test subject is denoted by S^t, then *FPAN* provides the final feature F_1 in which each attribute represents the probability of the test subject to belong to a particular class. Thus, we can write:

$$F_1 = \{F_1^1 F_1^2 ... F_1^N\}, \tag{4}$$

where,

$$F_1^j = \frac{|S^j - S^t|}{\sum_{j=1}^{N} |S^j - S^t|}, \forall j = 1, 2, ..., N. \tag{5}$$

Motion Pooling Network A Motion Pooling Network (*MPN*), with average pooling layers, has been employed as the second network, which preserves important information about the shape of a subject in each clip c by averaging the tensors f_c^t obtained corresponding to each frame t in the clip. Average pooling enables the preservation of useful dynamic information by aggregating clip-level temporal feature descriptors. It is understandable that, if the value of the clip length (i.e., the value of T) is *1*, the model will behave like a simple frame-based model, and will not be able to retain temporal features, while for higher values of T, the network will fail to capture the kinematic information of a person's movement at a high resolution. To determine the best configuration for this sub-network, we experimented with different values of T using the *MARS* data [58], which is an extensive data set with 1191003 images and 1262 identities (refer to Table 1), and observe that T=4 provides the best performance among all the different values of T considered here. Thus, for the *MPN*, the clip length T is set to *4*, and we use this configuration to report further results from other data sets as well. The tensor at the penultimate layer of this network is of size $[w, h, 2048]$, following which a fully connected classification layer is introduced with the number of nodes equal to the number of classes in the data set. This classification layer outputs a vector F_2 in which each attribute represents the probability of a test subject to belong to the corresponding class.

Long-Short Term Memory Network This network is used to capture recurrent information from a walking sequence. It is well-known that an *LSTM* network can represent any time-series data effectively. Since a walking sequence can also be looked upon as a time-series data, the features provided by the *LSTM* network are expected to preserve unique motion features for each subject. Specifically, we use two *LSTM* cells on top of the feature descriptor from the *ResNet-101* to generate correspondences among the frames in an input video

Table 1 Overview of the data sets used in the study, and the evaluation metric used

Data set names	Number of cameras	Number of images	Number of identities	Evaluation
PRID-2011 [17]	2	24541	749	CMC
iLIDS-VID [40]	2	42495	300	CMC
MARS [58]	6	1191003	1261	CMC/map

sequence. The inputs to the second *LSTM* cell are the original image frames along with the hidden layer features from the first *LSTM* cell. If the feature vector corresponding to clip c, as obtained from this second *LSTM* cell, is denoted by o_c, then

$$o_c = \sigma(W_o.[x_i, h_{i-1}] + b_o). \tag{6}$$

where W_o represents the weight matrix associated with the network, and x_i and h_{i-1} respectively represent the inputs at the current cell and output from the previous cell, and b_0 denotes the bias. Similar to *MPN*, here also we compute the averaged *LSTM* feature from all the clips and obtain the feature vector F_3 from the final layer whose attributes represent the individual class probabilities. To select the best *LSTM* configuration for this third sub-network, we test with different hidden state sizes (i.e., 256, 512, and 1024) using the *MARS* data set, and observed that the best hidden state size for this model is 512. The sequence length (T) used here is fixed to 8 to obtain effective temporal features.

Ensemble models have been previously used in [11, 35, 41, 50] to perform joint feature learning or to fuse predictions from multiple classifiers. Here, we also follow a similar ensemble-based approach by stacking the different deep models to make the final prediction about the class of a test subject based on the average probability obtained from the three sub-networks. Due to fusing information from the multiple models, our proposed approach results in accurate and reliable predictions. This stacked ensemble model has been named as Temporal Motion Aware Network (T-MAN) since it accumulates the prediction of several temporal motion models. For visualization, we also present the feature maps generated at the final convolution layer by each sub-network as well as the ensemble model *T-MAN* in Fig. 2a–d. Although, 256 different feature maps are computed at the penultimate layer of each sub-network, here we present randomly chosen 64 feature maps among these for ease of visualization.

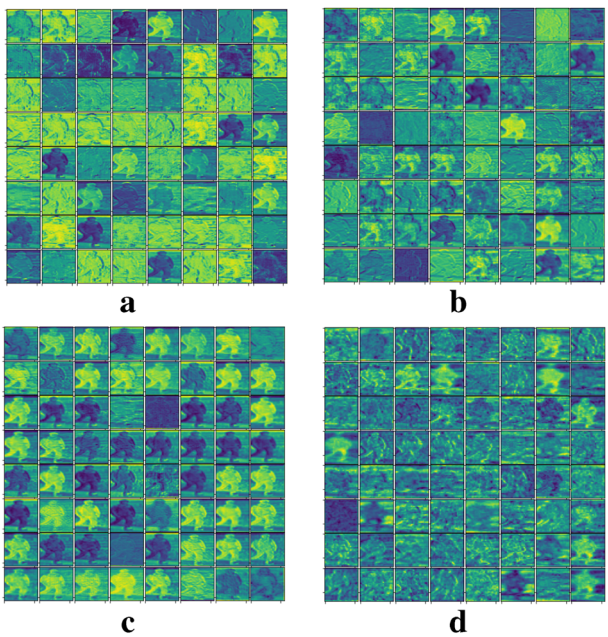

Fig. 2 Feature maps generated at the intermediate layers of **a** *FPAN*, **b** *MPN*, **c** *RNN*, **d** T-MAN

4 Evaluation settings

In this section, we briefly explain the details of the data sets used in the study, system configuration, and a thorough evaluation of our approach along with a comparative analysis with other competing approaches.

Data Sets The important characteristics (i.e., number of cameras, the total number of images, number of identities, and evaluation metric used) for each data set are highlighted in Table 1 along with relevant citations.

PRID-2011 [17]: This dataset consists of images from 749 persons captured by two non-overlapping cameras, and 200 individuals among these appear 749 subjects in both the camera views. The data set is less challenging since the images present here are captured in non-crowded regions with rare occlusion and relatively clear background. As in [33], for our experiments, we consider only the common set of *200* subjects that appear in the fields-of-view of both the cameras.

iLIDS-VID [40]: This data set consists of pedestrian images captured in an airport arrival hall. It is constructed from two non-overlapping camera views and contains 600 image sequences from 300 distinct individuals. This data set incorporates more challenging scenarios as compared to that in *PRID-2011* by considering occlusion, background clutter, viewpoint and lighting variations, etc. The number of frames present in the video sequences in this data set ranges from 23 to 192 with an average of 73.

MARS [58]: This data set is the largest video re-identification data set to date. It consists of about *20000* video sequences from *1261* individuals. Each of the sequences is obtained automatically by using the Deformable Part Model [12] detector. The tracking of individuals is carried out through the GMMCP [8] tracker. In this data set, video sequences of each person are captured by a minimum of two cameras and a maximum of six cameras. On average, it contains 13 video sequences for each person.

Evaluation Metrics Effectiveness of the proposed approach is evaluated by observing the rank-based recognition rate for different values of the rank. In most real-life applications of re-identification, finding the correct class as the best match of a classifier is not highly desirable. Rather, for practical purposes, it is sufficient if the correct class falls within the top few matches. Rank-based classification performance analysis provides a better estimate about the effectiveness of a classifier by providing not only the *Rank-1* accuracy but also accuracy values at some higher ranks. Apart from this, we also report the Mean Average Precision (map) of the proposed approach for the *MARS* data set and compare it with the state-of-the-art techniques.

Implementation Details We have implemented our algorithm using Python and Tensorflow [1] on a system having 64GB of RAM and NVIDIA TITAN Xp with GeForce GTX GPU with 34 GB memory capacity. The soft-max cross-entropy loss function has been used to train each of the individual networks, i.e., *FPAN*, *MPN*, and *LSTM*.

$$Loss_{(Softmax)} = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{a=1}^{C} g_{i,a}\ log\ p_{i,a}. \tag{7}$$

The training batch has been created by randomly selecting N identities and C clips for each identity. So, in total there are $(N \times C)$ clips in a batch, and let $g_{i,a}$ and $p_{i,a}$ denote the ground truth and prediction for sample *(i,a)*. We perform experiments with different

values of learning rate and weight decay, and observed that the values 0.0003 and 0.0005 suit best for the learning rate and weight decay factor, respectively. Training of each of the sub-networks, i.e., *FPAN*, *MPN*, *LSTM*, has been done for a maximum of 1000 epochs, or till the loss value does not alter significantly in two successive epochs.

5 Experimental evaluation

Here, we present the results obtained from the evaluation of the proposed T-MAN on different data sets as discussed in Section 4, and compare it with state-of-the-art re-identification techniques including [10, 14, 18–20, 24–26, 32, 40, 45, 47, 52, 55, 56]. As explained in Section 3, although *conv-LSTM* has been used as the final sub-network, to evaluate the effectiveness of the ensemble model, the combination of $FPAN$ and MPN has also been tested with other recurrent network models. In our first experiment, we perform this experiment by considering three different *RNN* cells, namely, (i) simple Recurrent Neural Network (RNN), (ii) Long-Short term Memory (*LSTM*) as in the proposed work, and (iii) Gated Recurrent Unit (*GRU*) using the data sets discussed in Section 4. The hidden state size (H_t) and sequence length (T) for each type of *RNN* cell have been fixed to 512 and 8, respectively. The results are shown in the form of a grouped bar chart in Fig. 3, in which each bar corresponds to a particular ensemble model as indicated in the legend of the plot. Data set names have been specified along the horizontal axis, whereas the vertical axis shows the *Rank-1* accuracy.

It can be seen from the figure that the proposed ensemble model *T-MAN* with *LSTM* as the recurrent layer performs better than any other ensemble model for all the data sets i.e., *PRID-2011*, *iLIDS-VID*, and *MARS*. This justifies the use of *LSTM* as a recurrent feature extractor over the other *RNN* models. In our next set of experiments, we compare the performance of the proposed approach with state-of-the-art techniques using each of the three data sets. Results for the first two data sets are shown in Tables 2 and 3 in terms of *Rank-1*, *Rank-5*, and *Rank-10* accuracy, whereas results for the third data set are shown in Table 4 in terms of *Rank-1*, *Rank-5*, and *Rank-20* accuracy and the *map* score. In each table, along with

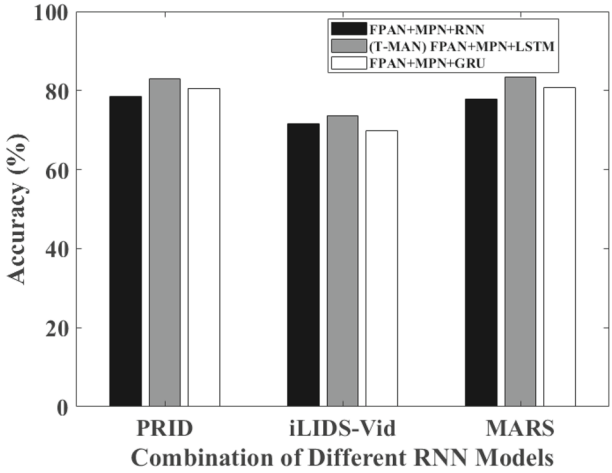

Fig. 3 *Rank-1* accuracy for different combinations of RNN Models (i.e., *Simple RNN*, *LSTM*, and *GRU*)

Table 2 Comparative results on *PRID-2011* data set for Ranks 1, 5 and 10

Data set	*PRID-2011*		
Methods	*Rank-1*	*Rank-5*	*Rank-10*
Baseline [33]	70.0	90.0	95.0
STA [26]	64.1	87.4	90.0
TDL [52]	56.7	80.1	87.6
DVR [40]	40.4	71.8	84.6
TAPR [14]	74.0	94.6	94.2
SRID [18]	35.2	59.5	70.0
RFA-Net [47]	58.2	85.8	93.5
AMOC [25]	83.7	98.3	99.4
TAUDL [23]	49.4	78.7	92.6
DGM+XQDA [49]	81.1	95.1	98.9
our *FPAN*	65.0	71.5	83.0
our *MPN*	62.4	72.0	85.5
our *LSTM*	81.5	88.2	93.5
our T-MAN (*FPAN+MPN+LSTM*)	83.0	96.4	98.8

the competing approaches, we have also tested the effectiveness of each of the sub-networks used in the proposed *T-MAN* model, if used separately for person re-identification.

From Table 2, it can be seen that for the *PRID-2011* data set, we have achieved quite satisfactory *Rank-1* accuracy of 83% with the proposed ensemble model T-MAN. For *Rank-10*, this accuracy is more than 98%, which can be said to be significantly good, given the data set consists of 749 identities. The benefits of choosing an ensemble model can be verified from the final four rows of this table. It can be seen that the ensemble model significantly improves upon the accuracy of the individual sub-networks, namely, *FPAN*, *MPN*, and *LSTM*. It is also observed from the table that although for this data set, the accuracy of our approach for the different rank values is significantly high, the approach in [25] performs slightly better than ours. Similar observation also follows from Table 3. Here, also our approach performs much better than most of the existing techniques. Only the results given by [25] are closely comparable with our work. We observe that although the *Rank-1* accuracy of our method exceeds that of [25] by about 5%, the *Rank-5* and *Rank-10* accuracy of [25] is slightly better than ours.

Table 4 shows a comparative performance analysis of our work with ensembles of other existing approaches on the *MARS* data set. Three different metric learning algorithms and seven feature descriptors have been used in this study, as given next. The descriptors include SDALF [10], HOG3D [19], HistLBP [45], gBiCov [32], LOMO [24], BoW [55] and IDE [56] whereas metric learning methods are DVR [40], KISSME [20], and XQDA [24]. Although among all the previous video-based re-identification approaches the work in [25] performs the best, the proposed *T-MAN*-based re-identification method outperforms this approach by 15% *Rank-1* accuracy. Thus, for large data sets, our ensemble model can be said to perform more reliably than that of [25]. Additionally, *T-MAN* has achieved a mean average precision (map) score of 76.7%, which is significantly higher than any of the state-of-the-art methods. The superior performance of the proposed T-MAN model is due

Table 3 Comparison results on *iLIDS-VID* data set for Ranks 1, 5 and 10

Data set	*iLIDS-VID*		
Methods	*Rank-1*	*Rank-5*	*Rank-10*
Baseline [33]	58.0	84.0	91.0
STA [26]	44.5	71.7	83.7
TDL [52]	56.5	87.6	95.6
DVR [40]	39.4	61.1	71.8
TAPR [14]	55.1	87.4	93.3
SRID [18]	25.0	44.5	55.6
RFA-Net [47]	49.3	76.7	85.4
AMOC [25]	68.7	94.3	98.3
TAUDL [23]	26.7	51.3	78.6
DGM+XQDA [49]	42.6	67.7	76.6
our *FPAN*	61.1	69.4	81.0
our *MPN*	59.0	66.5	76.4
our *LSTM*	64.9	77.2	85.0
our T-MAN (*FPAN+MPN+LSTM*)	73.5	91.4	96.6

to making prediction by fusing three important motion-related information from the three sub-networks.

Next, we observe the effectiveness of the different ensemble models that can be formed by combining two or more sub-networks used in the study, and report the

Table 4 Comparison results on *MARS* data set for Ranks 1, 5 and 20 with Mean Average Precision (map)

Data set	*MARS*			
Methods	*Rank-1*	*Rank-5*	*Rank-20*	Mean Average precision (*map*)
SDALF [10]+DVR [40]	4.1	12.2	25.4	1.8
HOG3D [19]+KISSME [20]	2.7	6.4	12.5	0.8
HistLBP [45]+XQDA [24]	18.2	33.1	46.0	8.0
BoW [55]+KISSME [20]	30.6	46.4	60.1	15.5
LOMO + XQDA [24]	30.8	46.4	60.9	16.5
gBiCov [32]+XQDA [24]	9.2	19.8	33.4	3.7
IDE [56]+XQDA [24]	65.5	82.0	89.0	47.5
AMOC+EpicFlow [25]	68.3	81.4	90.6	52.9
TAUDL [23]	43.8	59.9	72.8	29.1
DGM+IDE [49]	48.1	64.7	77.4	29.1
our *FPAN*	60.1	69.9	78.4	49.5
our *MPN*	57.5	66.1	79.0	45.0
our *LSTM*	68.7	79.8	89.2	53.4
our T-MAN (*FPAN+MPN+LSTM*)	83.3	93.5	95.6	76.7

Table 5 Comparative analysis of different combinations of proposed models (*FPAN*, *MPN*, and *T-MAN*)

Data set	*MARS*			
Methods	*Rank-1*	*Rank-5*	*Rank-20*	Mean Average precision (map)
Ensemble 1 (*FPAN+MPN*)	65.2	76.4	89.0	51.6
Ensemble 2 (*MPN+LSTM*)	73.1	89.0	92.6	64.9
Ensemble 3 (*FPAN+LSTM*)	79.2	92.1	95.5	74.1
our T-MAN (*FPAN+MPN+LSTM*)	83.3	93.5	95.6	76.7

Rank-1, *Rank-5*, and *Rank-20* accuracy for each in Table 5. Specifically, we consider the following ensemble models: Ensemble 1 (*FPAN+MPN*), Ensemble 2 (*MPN+LSTM*), and Ensemble 3 (*FPAN+LSTM*). The extensive *MARS* data set has been used for this experiment.

From this table also, we can see that the proposed ensemble model performs better than any other combination of the sub-networks.

In our final experiment, we evaluate the robustness of our proposed Temporal Motion Aware Network (T-MAN) against various initialization parameters of the three sub-networks. To do this, we first train the individual models three different times, and next ensemble these to get three different trained models. We test the performances of each of the above trained models and observe the *Rank-1* accuracy. Results are shown in the form of a grouped bar diagram in Fig. 4.

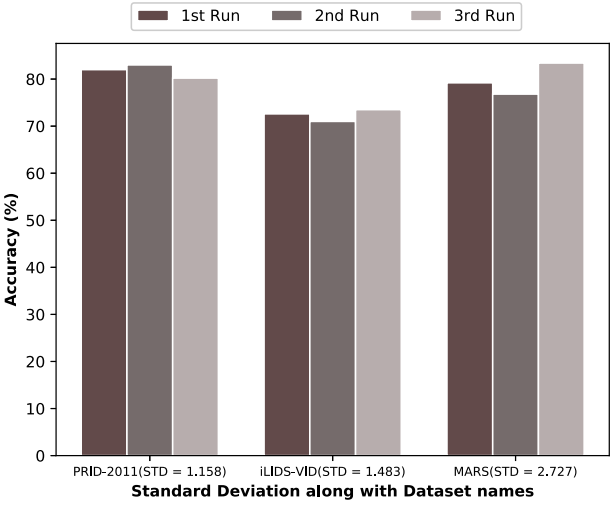

Fig. 4 *Rank-1* accuracy obtained by executing our ensemble T-MAN model three times along with the standard deviation

Here, the height of each bar corresponds to the *Rank-1* recognition accuracy, and each group of three bars represents the accuracy obtained by running the three different trained models on each data set. The data set names and the standard deviation of the recognition accuracy are shown along the horizontal axis. It is observed from the figure that the *Rank-1* accuracy for *PRID-2011*, *iLIDS-VID*, and *MARS* data set ranges between [80.2 83.0], [71.0 73.5], and [76.8 83.3], respectively. The standard deviation values for the *PRID-2011* and *iLIDS-VID* data sets are well below 1.5, which implies that the proposed T-MAN model is robust to varying initialization parameters of the sub-networks. A slightly higher value of standard deviation (i.e., 2.7) has been observed for the *MARS* data set. This is since the *MARS* data set consists of low resolution images that pose significant challenge to the classification algorithms. However, despite its slightly less robustness, the *Rank-1* accuracy values obtained from the differently trained models are remarkably high.

Discussions The above experimental results show that the proposed *T-MAN* based re-identification method is accurate, robust to varying initialization parameters, and also outperforms almost every state-of-the-art approach for the different experimental settings. We would especially like to emphasize the point that, during evaluation using the challenging *MARS* data set (which consists of very low resolution images), our technique performs better than the existing techniques by at least 15% in terms of *Rank-1* accuracy. Most of the previous approaches working on *MARS* data set have used either a single network model or ignored the important motion-related information from the video sequences. In contrast, we combine the contextual, motion, and temporal information into the Temporal Motion Aware Network (T-MAN) model to carry out person re-identification effectively.

6 Conclusions and future work

In this work, we propose an ensemble model *T-MAN* to perform video-based person re-identification by combining predictions from three different deep networks, namely, *FPAN*, *MPN*, and *LSTM*. The proposed model jointly handles temporal attention with motion and recurrent information from input video sequences. We comprehensively study and compare the performance of the proposed model in terms of rank-wise classification rate with state-of-the-art techniques. Our approach has been seen to outperform most of the state-of-the-art techniques for the different experimental settings used in the study, and it performs significantly better than the previous techniques for the extensive *MARS* data set, which consists of a large number of subjects. Effective spatio-temporal modelling and attention-based feature extraction are the main reasons behind the superiority of our model over the previous approaches.

Static and dynamic occlusion handling in crowded environments by employing *GAN*-based inpainting [3], or some recurrent network models, and open-set person re-identification maybe considered as future scopes for work. Instead of aggregating features from all the frames in a sequence, as done in the first two sub-networks, i.e., *FPAN* and *MPN*, pose-based feature extraction might help in improving the overall accuracy of the model, which maybe studied in the future. Also, saliency detection techniques, as described in [3, 42, 51], maybe incorporated in our model to extract the most important (i.e., salient) features at the frame level or the sequence level, and enable the sub-networks to make better predictions in a more time-efficient manner.

Acknowledgements The authors would like to acknowledge NVIDIA for supporting their research with a TITAN Xp Graphics processing unit.

References

1. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M et al (2016) Tensorflow: a system for large-scale machine learning. In: 12^{th} {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp 265–283
2. Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3908–3916
3. Cai W, Wei Z (2020) PiiGAN: generative adversarial networks for pluralistic image inpainting. IEEE Access 8:48451–48463
4. Chen L, Lou J, Xu F, Ren M (2019) Grid-based multi-object tracking with siamese CNN based appearance edge and access region mechanism. Multimed Tool Appl :1–19
5. Cheng DS, Cristani M, Stoppa M, Bazzani L, Murino V (2011) Custom pictorial structures for re-identification. In: Proceedings of the British machine vision conference. Citeseer, pp 1–11
6. Chung D, Tahboub K, Delp EJ (2017) A two stream siamese convolutional neural network for person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 1983–1991
7. Dai J, Li Y, He K, Sun J (2016) R-FCN: object detection via region-based fully convolutional networks. In: Proceedings of the advances in neural information processing systems, pp 379–387
8. Dehghan A, Modiri Assari S, Shah M (2015) GMMCP tracker: globally optimal generalized maximum multi clique problem for multiple object tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4091–4099
9. Fan D-P, Wang W, Cheng M-M, Shen J (2019) Shifting more attention to video salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8554–8564
10. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2360–2367
11. Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller P (2019) Deep neural network ensembles for time series classification. arXiv:1903.06602
12. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2009) Object Detection with Discriminatively Trained Part-Based Models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
13. Fu J, Zheng H, Mei T (2017) Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4438–4446
14. Gao C, Wang J, Liu L, Yu J-G, Sang N (2016) Temporally aligned pooling representation for video-based person re-identification. In: Proceedings of the IEEE international conference on image processing, IEEE, pp 4284–4288
15. Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings of the European conference on computer vision, Springer, pp 262–275
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
17. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Proceedings of the scandinavian conference on image analysis. Springer, pp 91–102
18. Karanam S, Li Y, Radke RJ (2015) Sparse re-id: block sparsity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 33–40
19. Klaser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of the 19^{th} british machine vision conference, pp 275:1–10
20. Koestinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H (2012) Large scale metric learning from equivalence constraints. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, pp 2288–2295
21. Kviatkovsky I, Adam A, Rivlin E (2012) Color invariants for person reidentification. IEEE Trans Pattern Anal Mach Intell 35(7):1622–1634

22. Li W, Zhao R, Xiao T, Wang X (2014) DeepReID: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 152–159
23. Li M, Zhu X, Gong S (2018) Unsupervised person re-identification by deep learning tracklet association. In: Proceedings of the European conference on computer vision, pp 737–753
24. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2197–2206
25. Liu H, Jie Z, Jayashree K, Qi M, Jiang J, Yan S, Feng J (2017) Video-based person re-identification with accumulative motion context. IEEE Trans Circuits Syst Video Technol 28(10):2788–2802
26. Liu K, Ma B, Zhang W, Huang R (2015) A spatio-temporal appearance representation for video-based pedestrian re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 3810–3818
27. Liu Y, Yan J, Ouyang W (2017) Quality aware network for set to set recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5790–5799
28. Lu X, Ma C, Ni B, Yang X (2019) Adaptive region proposal with channel regularization for robust object tracking. IEEE Trans Circuits Syst Video Technol
29. Lu X, Wang W, Ma C, Shen J, Shao L, Porikli F (2019) See more, know more: unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3623–3632
30. Ma M (2019) Infrared pedestrian detection algorithm based on multimedia image recombination and matrix restoration. Multimed Tools Appl :1–16
31. Ma B, Su Y, Jurie F (2012) Local descriptors encoded by fisher vectors for person re-identification. In: Proceedings of the European conference on computer vision. Springer, pp 413–422
32. Ma B, Su Y, Jurie F (2014) Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image Vis Comput 32(6-7):379–390
33. McLaughlin N, Martinez del Rincon J, Miller P (2016) Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1325–1334
34. Minetto R, Segundo MP, Sarkar S (2019) Hydra: An ensemble of convolutional neural networks for geospatial land Classification. IEEE Trans Geosci Remote Sens
35. Muñoz DU, Ruiz-Aguilar JJ, González-Enrique J, Domínguez IJT (2019) A deep ensemble neural network approach to improve predictions of container inspection volume. In: Proceedings of the international work-conference on artificial neural networks. Springer, pp 806–817
36. Nguyen HD, Na IS, Kim SH, Lee GS, Yang HJ, Choi JH (2019) Multiple human tracking in drone image. Multimed Tools Appl 78(4):4563–4577
37. Prosser BJ, Zheng W-S, Gong S, Xiang T, Mary Q (2010) Person re-identification by support vector ranking. In: Proceedings of the British machine vision conference, pp 1–11
38. Pytorch: Models. https://pytorch.org/docs/stable/torchvision/models.html. Accessed: 2020-01-16
39. Tagore NK, Singh SK (2019) Crowd counting in a highly congested scene using deep augmentation based convolutional network. Available at SSRN 3392307
40. Wang T, Gong S, Zhu X, Wang S (2016) Person re-identification by discriminative selection in video ranking. IEEE Trans Pattern Anal Mach Intell 38(12):2501–2514
41. Wang Z, Zou C, Cai W (2020) Small sample classification of hyperspectral remote sensing images based on sequential joint deeping learning model. IEEE Access 8:71353–71363
42. Wang Z, Zou C, Cai W (2020) Small sample classification of hyperspectral remote sensing images based on sequential joint deeping learning model. IEEE Access 8:71353–71363
43. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10(Feb):207–244
44. Wu Z, Shen C, Van Den Hengel A (2019) Wider or deeper: revisiting the resnet model for visual recognition. Pattern Recogn 90:119–133
45. Xiong F, Gou M, Camps O, Sznaier M (2014) Person re-identification using kernel-based metric learning methods. In: Proceedings of the European conference on computer vision, Springer, pp 1–16
46. Xu S, Cheng Y, Gu K, Yang Y, Chang S, Zhou P (2017) Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In: Proceedings of the IEEE international conference on computer vision, pp 4733–4742
47. Yan Y, Ni B, Song Z, Ma C, Yan Y, Yang X (2016) Person re-identification via recurrent feature aggregation. In: Proceedings of the European conference on computer vision, Springer, pp 701–716
48. Yang X, Chen P (2019) Person re-identification based on multi-scale convolutional network. Multimed Tools Appl :1–15

49. Ye M, Li J, Ma AJ, Zheng L, Yuen PC (2019) Dynamic graph co-matching for unsupervised video-based person re-identification. IEEE Trans Image Process 28(6):2976–2990
50. You H, Tian S, Yu L, Lv Y (2019) Pixel-level remote sensing image recognition based on bidirectional word vectors. IEEE Trans Geosci Remote Sens 58(2):1281–1293
51. You H, Tian S, Yu L, Lv Y (2019) Pixel-level remote sensing image recognition based on bidirectional word vectors. IEEE Trans Geosci Remote Sens 58(2):1281–1293
52. You J, Wu A, Li X, Zheng W-S (2016) Top-push video-based person re-identification. In: Proceedings of the ieee conference on computer vision and pattern recognition, pp 1345–1353
53. Zheng H, Fu J, Mei T, Luo J (2017) Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision, pp 5209–5217
54. Zheng W-S, Gong S, Xiang T (2012) Reidentification by relative distance comparison. IEEE Trans Patt Anal Mach Intell 35(3):653–668
55. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
56. Zheng L, Zhang H, Sun S, Chandraker M, Yang Y, Tian Q (2017) Person re-identification in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1367–1376
57. Zhou Z, Huang Y, Wang W, Wang L, Tan T (2017) See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4747–4756
58. Zhu X, Jing X-Y, You X, Zhang X, Zhang T (2018) Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. IEEE Trans Image Process 27(11):5683–5695

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.