Recurrent Convolutional Network for Video-based Person Re-Identification

Niall McLaughlin, Jesus Martinez del Rincon, Paul Miller
Centre for Secure Information Technologies (CSIT)
Queen's University Belfast
n.mclaughlin@qub.ac.uk

Abstract

In this paper we propose a novel recurrent neural network architecture for video-based person re-identification. Given the video sequence of a person, features are extracted from each frame using a convolutional neural network that incorporates a recurrent final layer, which allows information to flow between time-steps. The features from all time-steps are then combined using temporal pooling to give an overall appearance feature for the complete sequence. The convolutional network, recurrent layer, and temporal pooling layer, are jointly trained to act as a feature extractor for video-based re-identification using a Siamese network architecture. Our approach makes use of colour and optical flow information in order to capture appearance and motion information which is useful for video re-identification. Experiments are conduced on the iLIDS-VID and PRID-2011 datasets to show that this approach outperforms existing methods of video-based re-identification.

1. Introduction

The re-identification problem entails associating different tracks of a person as they move between non-overlapping cameras [7]. Accurate re-identification is crucial for robust wide-area tacking, where persons are tracked as they move through a camera-network, and may be useful for single-camera tracking, where short tracklets must be linked into longer more reliable tracks [24]. In the general case, person re-identification is difficult due to large appearance changes caused by environmental and geometric variations as a person moves between cameras.

In this work we address the problem of person re-identification in the video setting, which occurs when a video of a person as seen in one camera must be matched against a gallery of videos captured by a different non-overlapping camera. The problem of re-identification has been extensively explored for still images, however the video-based re-identification problem has not had the same attention, perhaps due to a lack of large video re-identification datasets in the past [43].

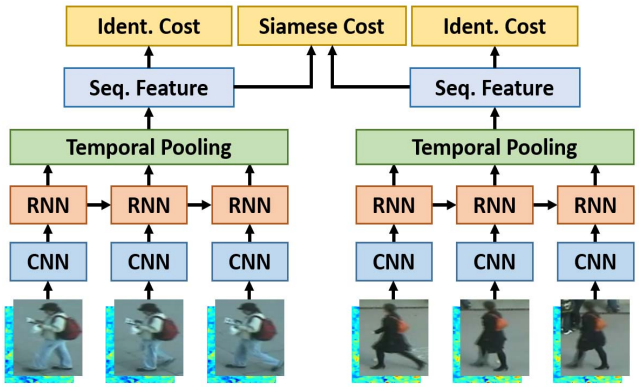

Figure 1. Our proposed video-based re-identification system.

The use of video for re-identification has several advantages over still images. The video setting is a more natural way to perform re-identification, as a person's image will normally be captured by a video camera, producing a sequence of images rather than a single still image. Given the availability of sequences of images, temporal information related to a person's motion, such as their gait, and perhaps even the way their clothing moves, is captured, which may help to disambiguate difficult cases that arise when trying to recognise a person in a different camera. Lastly, sequences of images provide a larger number of samples of a person's appearance, where each sample may have a different pose, viewpoint, and background, thus allowing a better model of the person's appearance to be built. The existence of a large number of samples makes it easier to train machine learning algorithms in general, and neural networks in particular. However, the use of video also creates several new challenges for re-identification, such as dealing with video sequences of arbitrary length and/or different frame-rates, the difficulty of creating an accurate appearance model given unknown partial or full occlusions within the sequences to be recognised, and the possibility of tracking inaccuracy that may arise when extracting the sequences. This final

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.148

problem is however mitigated by the emergence of accurate multi-target trackers [24].

2. Related work

Person re-identification for still images has been extensively studied with methods generally falling into two categories. The first of these employs invariant feature based methods that attempt to extract features that are both discriminative and invariant to environmental and view-point changes [23, 3, 6]. Secondly, supervised learning based methods that learn to map the raw features into a new space with greater discriminative power [14, 12, 44]. Deep learning techniques fall in this second category [45, 5, 8], and are deemed advantageous as they remove the need for hand-crafted features, and give improved performance provided there is sufficient training data. After features have been extracted, metric learning is widely used in person re-identification to learn a Mahalanobis metric that emphasises inter-personal distance and de-emphasises intra-person distance. The learnt metric is used to make the final decision as to whether a person has been on the re-identified or not. Various methods have been proposed based on this idea such as, Relaxed Pairwise Learning (PRLM) [14], Large Margin Nearest-Neighbour (LMNN) [44], and Relevance Component Analysis (RCA) [2].

While it is commonly assumed in many approaches to re-identification that each person is represented by a single image, the use of video in many realistic scenarios means that multiple images can be exploited to improve performance. Existing methods for multi-shot re-identification include collecting interest-point descriptors over time [9], or training classifiers using features collected over multiple frames [32]. In addition, supervised learning based methods have also been used, such as learning a distance preserving low-dimensional manifold [4], or learning to map between the appearances in sequences by taking into account the differences between specific camera pairs [25]. Other approaches that explicitly model video include using a conditional random field (CRF) to ensure similar images in a video sequence receive similar labels [18], or extracting space-time features [21, 1] and then learning a ranking function that is robust to partially corrupted sequences [43].

Recently, deep neural networks (DNN) have been successfully applied in many areas of computer-vision, such as large-scale object recognition [36, 22] and face recognition [35, 40], and in these areas they have largely replaced traditional computer vision pipelines based on hand-crafted features. In the area of image based person re-identification, DNNs have been used to learn ranking functions based on pairs [45], or triplets of images [5]. These methods, which use network architectures such as the 'Siamese network' [8], learn a direct mapping from the raw image pixels to a feature space where diverse images from the same person are close, while images from different persons are widely separated. Another DNN-based approach to re-identification, uses an auto-encoder to learn an invariant colour feature, whilst ignoring spatial features [42]. Specialised network architectures have been developed for directly comparing pairs of images taking into account deformation [26], which directly answer the question of whether two images depict the same person or not. Finally, several approaches have been proposed for improving generalisation given limited training data [30, 10]. However, existing architectures have been designed to represent spatial/appearance features but do not exploit any form of temporal information, and have not been applied to video re-identification before.

In order to introduce temporal signals into a DNN, architectural changes are required in conventional designs. Some attempts have been made in, for instance, action/event recognition from video, to understand features occurring over both the spatial and temporal dimensions with recurrent networks. These networks include feedback connections that allow the recall of events over time [33], and temporal-pooling networks, that average spatial features over multiple time-steps [34].

In this paper we propose a novel recurrent DNN architecture for video-based person re-identification. Our DNN-based system combines recurrency and temporal-pooling of appearance data with representation learning, by using a Siamese network architecture to learn an invariant representation for each person's video sequence. By introducing temporal pooling and recurrent layers, our proposed network architecture combines the data from all time-steps into a single feature vector for the whole input sequence, resulting in improved performance. This is the first time, to our knowledge, that deep learning has been applied to the video re-identification problem, which we consider to be the main contribution of this paper. Our proposed approach differs significantly from existing methods that are based on hand-crafted features, as it automatically learns to extract spatio-temporal features relevant for re-identification. Other important contributions of this work are: The use of temporal pooling to summarise the long-term appearance data of sequences with different lengths and frame-rates. The application of recurrency to emphasise temporal appearance data over the medium term. And finally, the use of both colour and optical flow pixel information as input to the DNN for re-identification, allowing it register short-term spatio-temporal information.

3. Method

A diagram of our proposed feature extraction architecture is shown in Fig. 1. In our architecture each frame is first processed by a convolutional neural network to produce a feature vector representing the person's appearance

at a particular instant in time. We then allow information to flow between time-steps by using a recurrent layer, before the outputs from all time-steps are combined using temporal pooling. Temporal pooling allows the network to summarise an arbitrarily long video sequence into a single feature vector, while the recurrent layer may allow the network to better exploit temporal information within the sequence, before the outputs from all time-steps are combined.

In order to train the feature extraction network to perform re-identification, we use a Siamese network architecture [8] as shown in Fig. 1. Given a pair of sequences from the same person, the Siamese architecture is trained to produce sequence feature vectors that are close in feature space, while given a pair of sequences from different persons, the network is trained to produce sequence feature vectors that are separated by a margin. This objective function mirrors the structure of the re-identification problem, where it must be decided whether two images depict the same person or not. In the following section we will explain each of the components of our proposed network in greater detail.

3.1. Input

The input to the convolutional network consists of both optical flow and colour channels. While colour encodes details of a person's appearance and clothing, optical flow directly encodes short-term motion, which may include details of a person's gait as well as other motion cues. By using both colour and optical-flow together, the network should be better able to exploit short-term temporal information in order to improve re-identification accuracy compared with using colour alone.

3.2. Convolutional Network

As shown in Fig. 1, at each time-step the image is processed by a convolutional neural network (CNN). The CNN involves many individual processing steps, therefore for notational simplicity we refer to the complete CNN as a function, $f = C(x)$, that takes an image x as input and produces a vector f as output. In general, a CNN processes an image using a series of layers, where each individual layer is composed of convolution, pooling, and non-linear activation-function steps. In our case, we use max-pooling and the hyperbolic-tangent (Tanh) activation-function. Each layer of the convolutional network therefore performs the operation $C'(s^{(t)}) = Tanh(Maxpool(Conv(s^{(t)})))$, where in the first layer, the input, $s^{(t)}$, is the original image, and in deeper layers the input is the output feature maps from the previous layer of the CNN.

Let $\mathbf{s} = s^{(1)}...s^{(T)}$ be a video sequence, of length T, consisting of whole-body images of a person, where $s^{(t)}$ is the image at time t. Each image, $s^{(t)}$, is passed through the CNN to produce a vector, $f^{(t)} = C(s^{(t)})$, where $f^{(t)}$ is the vectorised representation of the CNN's final layer activation maps. The vector $f^{(t)}$ is then passed forward to the recurrent layer (see Section 3.3), where it is projected into a low-dimensional feature-space and combined with information from previous time-steps. Note that the parameters of the CNN are shared across all time-steps meaning that each input frame is processed by the same feature-extraction network. Dropout [37] is used between the CNN and the recurrent layer in order to reduce over-fitting. Complete details of the CNN architecture are given Fig. 2.

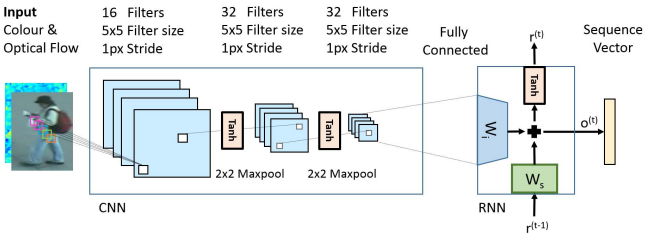

Figure 2. The structure of our proposed CNN and recurrent layer, where $r^{(t)}$ is the RNN's state at time t and $o^{(t)}$ is the sequence vector output at time t. See Section 3.2 and Section 3.3 for details.

3.3. Recurrent Layer

Recurrent neural networks (RNN) address the problem of processing an arbitrarily long time-series using a neural network, which can be problematic for standard architectures with a fixed number of input and output nodes. In contrast, a RNN has feedback connections, allowing it to remember information over time. At each time-step the RNN receives a new input and produces an output based on both the current input, and information from the previous time-steps. During training of a RNN using back-propagation-through-time, the recurrent connections are 'unrolled' to create a very deep feed-forward network [31], as shown in Fig. 1. Given the unrolled network, the lateral connections can be seen to act as memory, allowing information to flow between a potentially indefinite number of time-steps. It is commonly accepted that the performance of deep networks is due to hierarchical feature extraction that takes place over many layers [11], therefore we use a CNN to pre-process each input image into a higher-level representation before the recurrent layer.

As video re-identification involves recognising a person from a time-series of images, the use of recurrent connections may help to improve re-identification performance by allowing information to be passed between time-steps. By incorporating recurrent connections between the CNN and temporal pooling layers, we aim to better capture temporal information present in the video sequence.

As described in Section 3.2, $f^{(t)}$ is the vectorized output of the CNN's final layer activation maps, for the image

$s^{(t)}$ observed at time t. We can incorporate recurrent connections between the CNN and temporal-pooling layer as follows:

$$o^{(t)} = W_i f^{(t)} + W_s r^{(t-1)} \quad (1)$$

$$r^{(t)} = Tanh(o^{(t)}) \quad (2)$$

The output, $o^{(t)} \in \mathbb{R}^{e\times 1}$, at each time-step is a linear combination of the vectors, $f^{(t)} \in \mathbb{R}^{N\times 1}$, containing information on the current input image, and, $r^{(t-1)} \in \mathbb{R}^{e\times 1}$, containing information on the RNN's state at the previous time-step. The output is computed using the fully-connected layers, $W_i \in \mathbb{R}^{e\times N}$ and $W_s \in \mathbb{R}^{e\times e}$, respectively, where e is the dimensionality of the feature embedding-space, and N is the dimension of the vectorised representation of the CNN's final layer activation maps. Note that the parameter matrix W_i is non-square, meaning that the CNN's final-layer activation maps are projected to a vector in a lower-dimensional feature embedding space. The RNN state, $r^{(t)}$, is initialised to the zero-vector during the first time-step, $r^{(0)}$, and between time-steps is passed through the Tanh non-linear function.

3.4. Temporal Pooling

Although RNNs are able to capture temporal information, they have some drawbacks that may be relevant for re-identification. Firstly, the RNN's output may be biased towards later time-steps, making these more dominant than earlier ones [39, 15]. This could reduce the RNN's effectiveness when used to summarise the relevant information over a full sequence, because discriminative frames may appear anywhere in the sequence, not just near the end. Secondly, time-series analysis usually requires extracting information at different time scales. For instance in speech recognition, phonemes exist on a very short time-scale, and they are the building blocks for syllables, words, phrases, sentences, and conversations that exist at increasingly longer time scales. Since multiple time scales are not explicitly encoded in the standard RNN architecture, the temporal hierarchy present in the input signal may need to be explicitly embedded into the network design.

In order to address these limitations, our architecture adds a temporal pooling layer. This layer allows for the aggregation of information across all time steps, thus avoiding bias towards later time-steps. The temporal pooling layer aims to capture long-term information present in the sequence, which in combination with the short term scale of the optical flow input, and the middle-term recurrent layer, aims to model information at all temporal scales within the input signal.

In the temporal pooling layer, after forward propagation of a sequence of images, the appearance features produced by the combined CNN and recurrent layer for all time-steps, $\{o^{(1)} \dots o^{(T)}\}$, are aggregated to give a single feature representing the whole sequence. We propose two approaches to temporal pooling: In the first, mean-pooling is used over the temporal dimension to produce a single feature vector v representing the person's appearance averaged over the whole input sequence, as follows:

$$v_{\mathbf{s}} = \frac{1}{T}\sum_{t=1}^{T} o^{(t)} \quad (3)$$

In the second, max-pooling over the temporal dimension is used to select the maximum activation of each element of the appearance feature vector:

$$v_{\mathbf{s}}^{i} = \max([o^{(1),i}, o^{(2),i}, ..., o^{(T),i}]) \quad (4)$$

where $v_{\mathbf{s}}^{i}$ is the i'th element of the vector $v_{\mathbf{s}}$ and $[o^{(1),i}, o^{(2),i}, ..., o^{(T),i}]$ are i'th elements of the appearance vector across the temporal dimension. We now write the complete feature extraction network as a function $R(\mathbf{s}) = v_{\mathbf{s}}$, that takes as input a time-series of person images, $\mathbf{s}$, and produces a feature vector $v_{\mathbf{s}}$ as output, representing the person's appearance over the whole input sequence. This architecture allows sequences of arbitrary length to be compared by comparing each sequence's feature vector, rather than comparing the individual images at each time-step. In the following section we will explain how the above network can be trained to acts as a feature extractor, suitable for re-identification.

3.5. Training Strategy

3.5.1 Siamese Network

The proposed network can be trained to act as a feature extractor using the Siamese network architecture [8]. The Siamese network architecture consists of two sub-networks with identical weights. When the network is presented with a pair of inputs, the sub-networks map the pair of inputs to a pair of feature vectors, which are then compared using Euclidean distance. During training the Siamese network is shown similar and dissimilar input pairs, and it must learn to map those inputs to a feature space where similar inputs are close and dissimilar inputs are separated by a margin. Concretely, for video-based person re-identification we would like to map image-sequences from the same person to feature vectors that are close, and map sequences from different people to feature vectors that are widely separated.

Given a pair of sequences $(\mathbf{s}_i, \mathbf{s}_j)$, where each sequence has been processed using the feature extraction network to give sequence feature vectors, $v_i = R(\mathbf{s}_i)$ and $v_j = R(\mathbf{s}_j)$, we can write the Siamese network training objective as a function of the feature vectors v_i and v_j as follows:

$$E(v_i, v_j) = \begin{cases} \frac{1}{2}||v_i - v_j||^2 & i = j \\ \frac{1}{2}[\max(m - ||v_i - v_j||, 0)]^2 & i \neq j \end{cases} \quad (5)$$

where $||v_i - v_j||^2$ is the Euclidean distance between the feature vectors. When the sequences are from the same person i.e., $i = j$, the objective encourages the features v_i and v_j to be close, as measured by Euclidean distance, while for sequences from different persons i.e., $i \neq j$, the objective encourages the features to be separated by a margin m. During testing, features can be extracted for novel sequences, not observed during training, and whose identity is new and unknown, and these features can be compared using Euclidean distance, where a lower Euclidean distance indicates the sequences are more similar.

3.5.2 Joint Identification and Verification

Similar to the approach suggested in [38] for face recognition, we train the feature extraction network to satisfy both the Siamese objective and to predict the person's identity. Using the sequence feature vector, v, output by the feature extraction network, R, we can predict the identity of the person in the sequence using the standard cross-entropy loss, or softmax function, which is defined as follows:

$$I(v) = P(q = c|v) = \frac{\exp(W_c\, v)}{\sum_k \exp(W_k\, v)} \tag{6}$$

where there are a total of K identities, q is the identity of the person, and W_c and W_k refer to the c$^{\text{th}}$ and k$^{\text{th}}$ column of W, the softmax weight matrix, respectively. As an aside, we have found that jointly training for identification and Siamese cost is crucial for convergence. We can now define the overall training objective Q for a single pair of sequences, which jointly optimizes the Siamese cost and the identification cost as follows:

$$Q(\mathbf{s}_1, \mathbf{s}_2) = E(R(\mathbf{s}_1), R(\mathbf{s}_2)) + I(R(\mathbf{s}_1)) + I(R(\mathbf{s}_2)) \tag{7}$$

Where taking a similar approach to [38], we weight the identification cost and Siamese cost equally. The above network can be trained end-to-end using back-propagation-through-time (details of our training parameters can be found in section 4). During training with back propagation through time, all recurrent connections are unrolled to create a deep feed-forward graph, where the weights of the recurrent layer and CNN are shared between all time-steps [31]. After training we discard the Siamese and identification cost functions and retain $R()$ for use as a feature extractor, where the feature vectors extracted by $R()$ can be directly compared using Euclidean distance.

4. Experiments

In this section we evaluate our approach to video re-identification on two different datasets: iLIDS-VID [43] and PRID-2011 [12]. The iLIDS-VID dataset contains 300 persons, where each person is represented by two video sequences captured by non-overlapping cameras. The sequences range in length from 23 to 192 frames. The PRID-2011 dataset contains 749 persons, captured by two non-overlapping cameras, with sequences lengths of 5 to 675 frames. Following the protocol used in [43], we only consider the first 200 persons, who appear in both cameras.

For these experiments each dataset was randomly split into 50% of persons for training and 50% of persons for testing. All experiments were repeated 10 times with different test/train splits and the results averaged to ensure stable results. The hyper-parameters of the convolutional network were set to the same values as in [30], optimised for single-shot re-identification on the Viper re-identification dataset [7]. And based on [30], the margin in the Siamese cost function was set to 2, and the feature embedding-space dimension was set to 128. The network was trained for 500 epochs using stochastic gradient descent with a learning rate of 1e-3, and a batch size of one, alternating between showing the Siamese network positive and negative sequence pairs. A full epoch consisted of showing all positive sequence pairs and an equal number of negative pairs, random sampled from all training persons.

Given 150 persons with a maximum sequence length of 192 frames, training for 500 epochs takes approximately one day using an Nvidia GTX-980 GPU. Re-identification can then be performed efficiently, as only the new sequence must be passed through the network to produce a feature vector. Pre-computed feature vectors are stored for all gallery-sequences and can be very efficient compared with the new sequence using a single matrix vector product, in less than 1 second.

Positive and negative sequence pairs consist of two full sequences of arbitrary length from different cameras, showing the same person or different persons respectively. During training, sub-sequences of $k = 16$ consecutive frames were used for computational reasons, where a different subset of 16 consecutive frames over the full sequence length was randomly selected at each epoch. During testing we consider the first camera as the probe and the second camera as the gallery, as in [43].

Data augmentation in the form of cropping and mirroring was applied to increase the diversity of the training sequences, and for a given sequence the same augmentation was applied to all frames during each presentation to the network. During testing data augmentation was also applied to the probe and gallery sequences, and the similarity scores between sequences averaged over all the augmentation conditions, as in [16].

As a preprocessing step images were converted to the YUV colour space, before being passed to the network, and each colour channel was normalised to have zero mean and unit variance. Horizontal and vertical optical flow channels were calculated between each pair of frames using the

Lucas-Kanade algorithm [29]. The optical flow channels were then normalised to fall within the range -1 to 1. When training and testing with both optical flow and colour information, the first layer of the neural network used five input channels, three for colour and two for optical flow, and when training and testing with colour information only, three input channels were used.

4.1. Feature Type and Recurrent Connections

In this experiment we investigate some of the main architectural choices of our proposed system: the use of recurrent connections, and the choice of input channels. Training and testing of the network was performed with recurrent connections either disabled or enabled, and with either colour features only, or colour and optical flow features together. The results of this experiment are presented in Fig. 3 as CMC curves for the iLIDS-VID and PRID-2011 datasets.

The results show that the use of recurrent connections improves performance on both datasets regardless of the features types used, compared to the network without recurrent connections. For both datasets the best performance occurs when recurrent connections are enabled, and optical flow and colour features are used together. Performance is lowest for both datasets when recurrent connections are disabled and colour features are used alone. This suggests that our choice to explicitly embedded short term and medium term temporal information into the network architecture through the use of optical flow and a recurrent layer respectively, improves re-identification performance. For the iLIDS-VID dataset this benefit is more obvious, as there is a clear separation between the performance of different methods, while for PRID-2011 dataset the performance tends to be similar, as well as very high, after rank five. Qualitative examination of the data suggests that the iLIDS-VID dataset has more cluttered backgrounds and occlusion, showing a higher complexity than PRID-2011, where the subjects are more distinct. This lower complexity may explain why all variants of our proposed method perform similarly on the PRID-2011 dataset after the candidates with similar appearance,who are more likely to be confused, are grouped together in the first five ranks and upwards.

4.2. Temporal Pooling

In section 3.4 we proposed two methods for temporal-pooling of appearance information over a sequence to give a representation of the sequence as a single feature vector: mean-pooling and max-pooling.

In this experiment we compare re-identification performance when the network has been trained and tested with either mean-pooling or max-pooling, and with the recurrent connections disabled to make the effect of the different pooling methods clearer. We also consider a baseline method [30] for computing a similarity-score between sequences that processes each frame individually using a single-frame CNN trained using a Siamese architecture and whose individual frame outputs are combined into a single decision without mean-pooling: The similarity between the sequences is then taken as the average Euclidean distance between corresponding frames. This single-shot CNN is exposed to all the data from the video sequences available in training, and trained using pairs of still-images, rather than sequence pairs, where a different single frame over the full sequence length was randomly selected at each epoc. In this experiment training and testing was carried out using the iLIDS-VID dataset.

The CMC curves of the two pooling methods and the baseline approach are shown in Fig. 4. It can be seen that mean-pooling performs better than both max-pooling and the baseline method. These results are interesting as they show that using mean-pooling to represent the whole sequence as a single feature vector leads to better performance than the baseline method which considers each frame individually. They also shows the utility of considering all the time steps equally important in the decision by using mean pooling, as opposed to max-pooling where only the feature value in the temporal step with the largest activation is employed. These results suggest that using mean-pooling over the temporal sequence of features may allow the network to better cope with noise and/or occlusions, and produces a single robust feature vector to compress and represent the person's appearance over a period of time.

4.3. Probe and Gallery Sequence Lengths

It is reasonable to assume that the availability of more samples for each person will improve re-identification accuracy, however, the rate at which performance increases in relation to the availability of samples is unclear. In this experiment we investigate how re-identification accuracy varies depending on the lengths of the probe and gallery sequences during the test phase, assuming a pre-trained network. Testing was performed on the iLIDS-VID dataset, and the lengths of the probe and gallery sequences were varied between 1 and 128 frames, in steps corresponding with the powers-of-two. Training lengths were fixed to 16 time steps as indicated at the start of this section. For some cases, where the desired gallery or probe length is greater than the real sequence length, we simply use the whole sequence. Probe sequences of length k are taken from the first k frames of the sequence recorded by first camera, and the gallery sequences of length k are taken from the last k frames of the sequence recorded by the second camera, since those are the farther temporal instants respectively. Results are reported in Fig. 5 as a matrix showing the rank 1 re-identification accuracy as a function of the probe and gallery sequence lengths.

The results show that increasing either the probe or

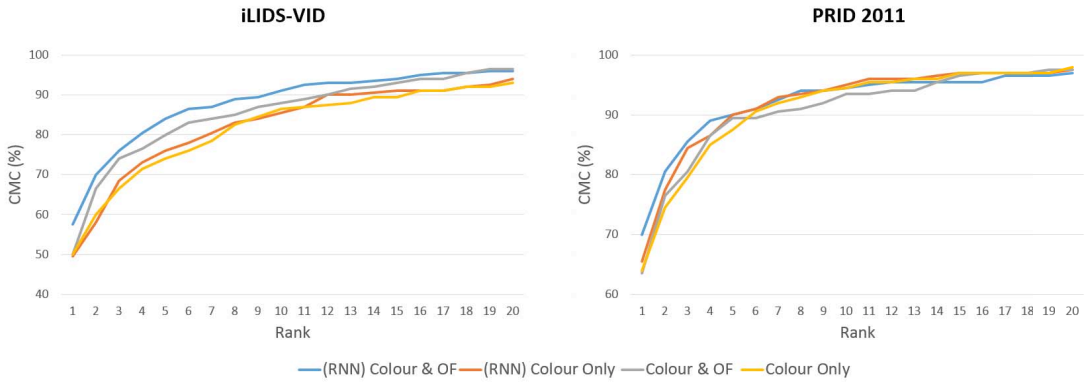

Figure 3. CMC curves for iLIDS-VID and PRID-2011 datasets, comparing the network trained and tested on with/without recurrent connections, and with colour and optical flow input, or colour input only. Note, the vertical axis in each figure have different scales.

gallery sequence lengths improves re-identification accuracy, and increasing both simultaneously gives the greatest improvement in accuracy, as can be noticed by the increasing CMC values in the diagonal. When different sample lengths are used, there seems to be approximate symmetry in performance when increasing either the probe sequence length or the gallery sequence length, with a slight benefit to having longer gallery sequences than probe sequences. This could prove useful for practical applications where it may be easier to collect large amounts of gallery data but where only a short probe sequence is available. When only one sample is available for each person in the gallery, increasing the probe length does not significantly improve accuracy, while if only one sample is available for the probe, increasing gallery length has a much greater effect on accuracy. This is of particular interest for those applications, such as watch-lists, where image to video re-identification is desired.

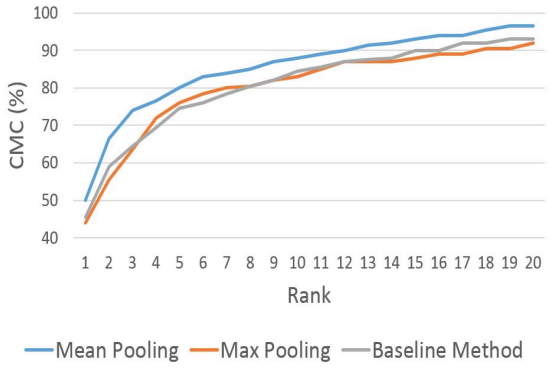

Figure 4. CMC curves comparing different methods of computing the similarity between sequences. Two temporal pooling architectures, mean-pooling and max-pooling, are compared with a baseline method without temporal pooling.

Gallery Sequence Length

Probe Sequence Length	1	2	4	8	16	32	64	128
1	14	19	21	22	23	26	25	27
2	15	20	22	22	26	26	29	31
4	14	20	26	23	28	30	31	33
8	15	23	25	28	31	34	38	41
16	19	24	30	31	36	41	43	43
32	20	26	33	33	39	44	47	46
64	19	28	32	33	38	44	50	51
128	18	27	33	35	40	45	52	52

Figure 5. iLIDS-VID rank 1 CMC re-identification accuracy as the lengths of the probe and gallery sequences are varied.

4.4. Comparison with the state of the art

We now compare the performance of our proposed video-based re-identification system against state-of-art methods from the literature. We also include results for the baseline DNN [30], described in Section 4.2, to put our results in context and to measure the improvement when using temporal information, as in our proposed network architecture. To ensure a fair comparison, the baseline system was trained and tested using the same datasets and same test/training split as the video-based system.

In Table 1 we compare the CMC results for our system, trained and tested on the iLIDS-VID and PRID-2011 datasets, with other state-of-the-art video re-identification systems. Comparing the CMC results of our proposed system with the baseline (still image based) system we can see that the video re-identification system performs better for both datasets. When we compare our results with the literature, our system shows superior performance against other video re-identification systems. The fact that even the baseline system shows better results than the existing state-of-the-art methods, shows the utility of DNNs in the re-identification context, as has been demonstrated in

many other application fields where sufficient training data is available [22, 36].

Dataset	PRID-2011				iLIDS-VID			
CMC Rank	1	5	10	20	1	5	10	20
Ours	70	90	95	97	58	84	91	96
Baseline	55	85	94	97	38	62	71	79
STA [28]	64	87	90	92	44	72	84	92
VR [43]	42	65	78	89	35	57	68	78
SRID [19]	35	59	70	80	25	45	56	66
AFDA [27]	43	73	85	92	38	63	73	82
DTDL [20]	41	70	78	86	26	48	57	69

Table 1. Comparison of our proposed approach with the literature on iLIDS-VID and PRID-2011 in terms of Rank CMC (%).

4.5. Cross-Dataset Testing

Cross-dataset testing may be a better way to estimate a system's real-world performance than evaluating performance on the same dataset used for training, which may lead to overfitting to a particular scenario. This is due to dataset bias [30, 41], which is a form of over-fitting where the performance of a machine-learning based system, trained on a particular dataset, is much worse when evaluated on a different dataset. One cause of this problem is that any given dataset represents only a small fraction of all real-world data, making it difficult for the system to learn which aspects of the training data are essential to the problem, and which are just artefacts of the dataset.

System	Trained On	1	5	10	20
Ours	iLIDS-VID	28	57	69	81
Ours*	iLIDS-VID	14	38	51	70
Baseline	Viper	17	36	48	68
Baseline*	Viper	14	31	45	61
CD [17]*	Shinpuhkan 2014	17	-	43	52

Table 2. Cross-dataset testing accuracy tested on PRID 2011 in terms of Rank CMC (%), where * indicates only one image was used for gallery and probe i.e. single-shot re-identification.

Therefore to better understand how well our proposed system generalises, we also perform cross-dataset testing, where the large and diverse iLIDS-VID dataset was used for training, and testing was performed on 50% of the PRID 2011 dataset, so that the results of this experiment can be compared with the results in Section 4.4 . We also include results for the baseline system comparison trained on the Viper dataset (for details of the baseline system please see Section 4.2). Testing was performed either using both the full sequences available, and to facilitate fair comparison with the literature, using a single still-image for both the probe and gallery for each person.

We can compare the results in the cross-dataset scenario with those in Table 2, when the system was trained and tested on PRID 2011 dataset. The results in the cross-dataset scenario are worse, as expected, probably due to dataset bias. However it should be noted that the rank 1 performance is not much below [19] (see Table 1), and is well above other single-shot re-identification systems, such as [17], even those specifically trained in PRID, such as [13] with a rank 1 CMC scores of 28. It can also be noticed there is a 100% improvement when using video re-identification that includes temporal information, which shows that our architecture is exploiting this temporal information to achieve better performance than the baseline. We include these results in the hope that others will also perform cross-dataset testing and improve the generalisation performance of re-identification systems.

5. Conclusion

In this paper we have introduced a novel temporal deep neural network architecture for video-re-identification. The use of optical flow, recurrent layers and mean-pooling allows us to embed the temporal hierarchy inherent to the problem in the form of short, middle and long term temporal information respectively. Results were evaluated in two standard datasets, and surpass any other method in the video re-identification literature. As future work, we plan to combine the current methodology with real multi target tracking outputs. This will make it possible to evaluate the robustness of our proposal when more noisy, fragmented and corrupt sequences are used as input, as well as to validate its applicability as a component of a full integrated wide area tracking system.

References

[1] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot human re-identification by mean riemannian covariance grid. In *AVSS*, pages 179–184. IEEE, 2011. 2

[2] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*, 6(6):937–965, 2005. 2

[3] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, volume 1, page 6, 2011. 2

[4] D. N. T. Cong, C. Achard, L. Khoudour, and L. Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In *ICIAP*, pages 179–189. 2009. 2

[5] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015. 2

[6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010. 2

[7] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, volume 3, 2007. 1, 5

[8] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742, 2006. 2, 3, 4

[9] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ICDSC 2008*, pages 1–6, 2008. 2

[10] D. Held, S. Thrun, and S. Savarese. Deep learning for single-view instance recognition. *arXiv preprint arXiv:1507.08286*, 2015. 2

[11] M. Hermans and B. Schrauwen. Training and analysing deep recurrent neural networks. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 190–198. 2013. 3

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Image Analysis*, pages 91–102. 2011. 2, 5

[13] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011. 8

[14] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793. 2012. 2

[15] S. Hochreiter and J. Schmidhuber. Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, pages 473–479, 1997. 4

[16] A. G. Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013. 5

[17] Y. Hu, D. Yi, S. Liao, Z. Lei, and S. Z. Li. Cross dataset person re-identification. In *ACCV Workshops*, pages 650–664, 2014. 8

[18] S. Karaman and A. D. Bagdanov. Identity inference: generalizing person re-identification scenarios. In *ECCV Workshops*, pages 443–452, 2012. 2

[19] S. Karanam, Y. Li, and R. Radke. Sparse re-id: Block sparsity for person re-identification. In *CVPR Workshops*, pages 33–40, 2015. 8

[20] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *ICCV*, 2015. 8

[21] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, pages 275–1, 2008. 2

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 1097–1105. 2012. 2, 8

[23] I. Kviatkovsky, A. Adam, and E. Rivlin. Color invariants for person reidentification. *Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013. 2

[24] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. MOTChallenge 2015: Towards a benchmark for multi-target tracking. *arXiv:1504.01942 [cs]*, Apr 2015. arXiv: 1504.01942. 1, 2

[25] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601. IEEE, 2013. 2

[26] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 2

[27] Y. Li, Z. Wu, S. Karanam, and R. J. Radke. Multi-shot human re-identification using adaptive fisher discriminant analysis. In *BMVC*, 2015. 8

[28] K. Liu, B. Ma, W. Zhang, and R. Huang. A spatio-temporal appearance representation for video-based pedestrian re-identification. In *CVPR*, pages 3810–3818, 2015. 8

[29] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981. 6

[30] N. McLaughlin, J. Martinez-del Rincon, and P. Miller. Data-augmentation for reducing dataset bias in person re-identification. pages 1–6, Aug 2015. 2, 5, 6, 7, 8

[31] M. C. Mozer. A focused back-propagation algorithm for temporal pattern recognition. *Complex systems*, 3(4):349–381, 1989. 3, 5

[32] C. Nakajima, M. Pontil, B. Heisele, and T. Poggio. Full-body person recognition system. *Pattern recognition*, 36(9):1997–2006, 2003. 2

[33] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. *arXiv preprint arXiv:1503.08909*, 2015. 2

[34] L. Pigou, A. v. d. Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *arXiv preprint arXiv:1506.01911*, 2015. 2

[35] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. *arXiv preprint arXiv:1503.03832*, 2015. 2

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 8

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 3

[38] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014. 5

[39] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 4

[40] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708, 2014. 2

[41] A. Torralba, A. Efros, et al. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011. 8

[42] R. R. Varior, G. Wang, and J. Lu. Learning invariant color features for person re-identification. *arXiv preprint arXiv:1410.1035*, 2014. 2

[43] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703. 2014. 1, 2, 5, 8

[44] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009. 2

[45] D. Yi, Z. Lei, and S. Z. Li. Deep metric learning for practical person re-identification. *arXiv preprint arXiv:1407.4979*, 2014. 2