Computing and Informatics, Vol. 41, 2022, 905–930, doi: 10.31577/cai_2022_4_905

NOVEL ARCHITECTURE FOR HUMAN RE-IDENTIFICATION WITH A TWO-STREAM NEURAL NETWORK AND ATTENTION MECHANISM

Babak Rahi

Department of Electronics and Computer Engineering
Brunel University London, Uxbridge, UK
e-mail: babak.hrahi@brunel.ac.uk

Man Qi

School of Engineering, Technology and Design
Canterbury Christ Church University, Canterbury, UK
e-mail: man.qi@canterbury.ac.uk

Abstract. This paper proposes a novel architecture that utilises an attention mechanism in conjunction with multi-stream convolutional neural networks (CNN) to obtain high accuracy in human re-identification (Reid). The proposed architecture consists of four blocks. First, the pre-processing block prepares the input data and feeds it into a spatial-temporal two-stream CNN (STC) with two fusion points that extract the spatial-temporal features. Next, the spatial-temporal attentional LSTM block (STA) automatically fine-tunes the extracted features and assigns weight to the more critical frames in the video sequence by using an attention mechanism. Extensive experiments on four of the most popular datasets support our architecture. Finally, the results are compared with the state of the art, which shows the superiority of this approach.

Keywords: Identification of persons, multi-layer neural network, gait recognition, human re-identification, convolutional neural networks, attention mechanism

1 INTRODUCTION

Gait recognition addresses the problem of human re-identification (Reid) at a distance by identifying people based on their motions and movements and how they walk using no overlapping cameras. Most human re-identification methods focus on the biometric features, including iris and face and are image-based. The problem of gait re-identification is better tackled through video analysis since it contains spatial and temporal information. In recent years, the study of Gait Data attracted the attention of scientists in different fields. Moreover, automated person re-identification (Reid) from large quantities of video surveillance data from non-overlapping cameras seems essential to the safety and security of our future generations. In past years, several models have been suggested for person Reid [1] and even calculation of the severity of diseases like Parkinson [2]. Gait recognition addresses the problem of human re-identification at a distance by identifying people based on their motions and movements (the way they walk). The gait recognition system is an unobtrusive biometric feature, which has attracted many researchers in recent years. Human motion analysis with visual tools is the attempt to 1. Detect, 2. Track and 3. Identify people, understand their behaviour and ideally predict their intentions from a series of image sequences that usually come from the frames extracted from a video [3, 4].

Over the past few years, researchers tried to combine motion analysis concepts and biometrics technology in surveillance systems. Recently computer vision community has shown immense interest in vision-based human re-identification at a distance. This interest is driven by the need to automate visual surveillance and monitoring systems for security-sensitive areas like banks, tube stations, parking lots, shopping centres and airports. The complexity of the gait recognition problem results from factors that can affect the accuracy of the recognition. These factors include – but are not limited to – multiple camera views, clothes, physical conditions, or even carrying a backpack. It is worth mentioning that the purpose of this work is not to build a strong classifier that can detect people only in a closed dataset but to build a system that can extract features from people outside of the dataset and re-identify them using this method. This model should be applicable in different situations and for different datasets, thus solving the problem of transfer learning since the popular datasets are relatively small ones. There are three approaches to gait feature extraction:

1. model-based features,
2. model-free features that are also called handcrafted features and
3. deep learned features.

Model-based approaches obtain a series of static or dynamic body parameters via modelling or tracking body components such as limbs, legs, arms and thighs. Gait signatures derived from these model parameters are employed to identify and recognise an individual: [4, 5, 6, 7] present some of the classic model-based approaches.

In a model-free approach, different features are extracted, like the whole motion of human bodies and silhouette width vector of Fourier descriptors. It also focuses on silhouette shape and the dynamic information used for pattern matching. To have an effective human re-identification model, we must identify discriminative features at high and low semantic levels. Recent models either use one semantic level feature representation or annotate these factors in the hardest possible way and use it in a single CNN network [6]. Some even use a Multilevel Factorization Net (MLFN), which uses no manual annotation [8]. A state-of-the-art approach on the use of CNN models for gait recognition was introduced in [9], however. It does not incorporate useful spatial features. We built on the work done in this paper to boost accuracy and the training time achieved a little bit further. We justify using optical flow maps by assuming that they hold all the crucial temporal information about the gait, arm swing, leg swing and velocity of motion. The spatial features include appearance characteristics such as colour, shape, size, and clothing. Furthermore, in low-resolution videos with insufficient pixel information for other biometric identification forms, gait recognition using optical flow shows excellent promise [10] since only the movement of human body points is considered.

We propose a deep spatial-temporal architecture comprising a pre-processing block, an STC block that contains a two-stream CNN with multiple fusions and an STA block that incorporates long short-term memory to focus the features extracted by the STC block. Our approach is not based on pose estimation or silhouette masks to get a motion model. Instead, it is based on training a convolutional neural network (CNN) and using the features extracted during the training of our network. Our contributions in this paper are as follows:

- A STC block with a novel two-stream spatial-temporal convolutional neural network with two concatenation fusions where we extracted the best spatial and temporal features simultaneously and achieved a more efficient exploitation of the available labelled data by proposing a new approach based on spatial-temporal architecture with multiple modality fusion for gait feature extraction which uses the best network modality combination for gait recognition.

- A novel STA block with a spatial-temporal attention mechanism with LSTM comprised a spatial and a temporal network. Improving the efficiency of the training stage for gait feature extraction which extracts the discriminative sequence-level features for representing the periodic motion cues of irregular gait sequences. The designed dual spatial attention mechanism can concentrate on the discriminative identity-related semantic regions from the spatial feature maps. The proposed mechanism for temporal attention can automatically assign adaptive weights (attention) to enhance the discriminative gait timesteps and suppress the redundant ones.

- Use of velocity of movement in the attention mechanism by incorporating this into the temporal attention weight.

- Extensive comparison between twenty combinations of the proposed approach and state of the art on four publicly available datasets, PRID 2011, iLIDS-VID, TUM-GAIT and CASIA.

This paper is structured in the following way. First, we reviewed the related work in Section 2. Then, in Section 3, we describe our approach in details. Next, Section 4 shows how we designed and conducted the four datasets and introduces the datasets. Finally, Section 5 discusses the conclusions made and possible future work.

2 RELATED WORK

There are numerous different approaches to gait recognition, most of which focus on extracting features from silhouettes, for example, based on Gait Energy Image (GEI) or pose estimation. Some only use binary silhouettes like [10]. Others like [11] propose a Patch Distribution Feature (PDF). They present each Gait Energy Image (GEI) as a set of local augmented Gabor features in concatenation with Gabor features extracted from different scales and orientations (40 D Gabor features extracted from 5 different scales and eight different orientations) together with a 2D x-y coordinate. The authors in [12] proposed a combination of Enhanced Gabor (EG) and representation of GEI and Regularised Locally Tensor Discriminant Analysis (RLTDA).

Furthermore, some approaches use mixed measurements of a subject's body like relational joint distance similar to [13], height, the shadow length in each stride, or even skeletal data [14] to generate an inner gait model based on computable values such as their standard deviation and mean. These approaches then select features and classify based on those nominated descriptors. Person re-identification for images has been the area of study for years. The two main focuses are on unsupervised learning (1), which employs methods that take invariant features to the environment, background, and viewpoint changes [15, 16]. Supervised learning (2), which learns to map the features into new spaces (feature maps) to increase the accuracy of identification and use traditional classification [17, 18]. Moreover, the deep learning techniques that fall into the category of supervised learning have the disadvantage of being reliant on the training data but do not need handcrafted features and also improve performance to a noticeable level [19, 20, 21]. After feature extraction, the distance between two points in terms of standard deviation (Mahalanobis distance) is learnt by the network and is used in the classification [22, 23]. Deep Neural Networks have been successfully applied to different problems in the past years, but they usually benefit from the spatial features only and do not focus on the temporal information [24, 25, 26, 27].

Especially in gait person re-identification, temporal information is a vast resource that should not be neglected. In recent years CNNs have been used to extract and use temporal features in human re-identification. [28, 29, 30] show the power of recurrent neural networks (RNNs) through feedback connections that allow an event

to be repeated over a period and temporal pooling that take an average of the spatial steps over a period of time. [31] proposes a recurrent deep neural network, which combines recurrence and temporal pooling with representation learning and learns invariant representation for each person of interest. This network converts the data from all time-steps into a feature vector which will help boost classification accuracy. The output from the CNN is fed into the RNN using a fully connected layer where the output of the RNN, $o^{(t)}$, is the feature vector at time t and $r^{(t)}$ is the state of the network at time t. We will expand on this more in Section 3 of this paper. Also, [32] presented an end-to-end architecture that combines CNNs with RNNs in a way that features of each frame are extracted using CNN and then fed into the RNN to get a complete spatial-temporal representation of the video.

Deep learning approaches based on CNNs have been used in image-based tasks with great success [33]. In the past years, deep architectures for video have appeared, and they are primarily focused on action recognition, where the inputs of the CNN are sub-sequences of stacked frames. No handcrafted features are used in these approaches, and all features are trained inside the neural network. [34] proposed to use as input to a CNN a volume obtained as the concatenation of frames with two channels that contain the optical flow in the x-axis and y-axis, respectively. They broke the CNN into two parts, a spatial stream convNet that gets video frames and a temporal stream convNet with optical flow maps as inputs. Each stream is composed of five convolutional layers, two full layers and one SoftMax layer and the output of these two are fed to an outside classification block. Other authors, such as [9], built on the idea from [34] to feed optical flow into the CNN with great results, but the performance was low for practical application. In the past, optical flow proved a helpful tool only for facial expression recognition by making feature point tracking easier [35]. However, their proposed pipeline computes the optical flow for the entire sequence, then builds up a cuboid by cropping and stacking the optical flow maps and feeds the cuboid into the CNN to output a unique gait signature in the form of a vector. It then applies a classifier to identify the subject. Our method is based on [34] and [9] fused with an attention network to boost the accuracy and performance.

In recent years more approaches to gait recognition have been proposed using millimetre wave sensing [36] or accelerometers [37] with hidden Markov models, but these are out of the scope of this paper.

3 PROPOSED APPROACH

Our approach is based upon a two-stream network presented for the first time in [34], which incorporates spatial and temporal networks and uses optical flow to attain high accuracy with limited training data and also uses multitask learning. However, their method has two significant disadvantages. Firstly, it uses only classification scores to perform fusion on the spatial and temporal features, and, as a result, it cannot learn some features, including pixel-wise correspondences. Secondly, it

is restricted since the spatial CNN only works on one frame and the temporal on limited optical flow frames. In our approach, we have developed the networks for the spatial-temporal features, address the fusion problem, and then introduce RNN to the mix to vectorise the results into one gait signature to boost the classification accuracy. Our input to the network is almost the same as the one proposed in [9, 31]. The output represented raw images (some pre-processing or zero padding will take place as part of the architecture for different datasets) combined with optical flow information for getting encoded details of a person of interest like appearance, gender and clothing, and their motion cues and gait details. Figure 1 represents our proposed architecture, which was implemented in this work. It shows an input block that crops and resizes images based on datasets used in the experiment to feed into the spatial stream of the STC block and the optical flow generator to use in the temporal stream of the STC block.

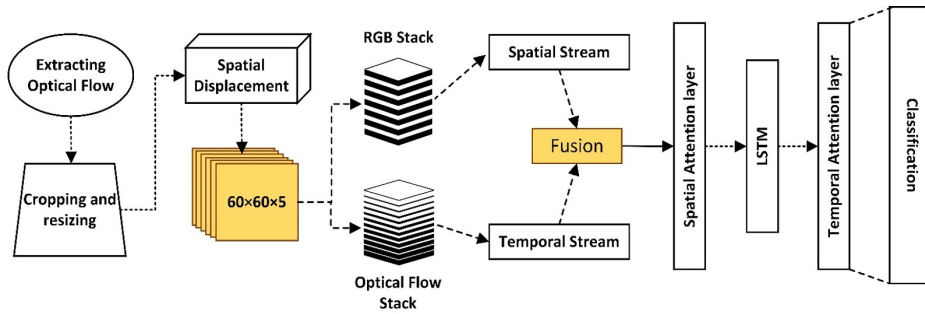

Figure 1. The illustration of our proposed architecture

The use of optical flow in action recognition has been proven to be an effective measure in [34], but it is possible to use optical flow to extract the well-defined and local features from the human gait. Simply put, the cuboid is made by stacking (concatenating) optical flow maps together to make a block fit for feeding to our temporal CNN. This form of input makes the recognition easier since it allows for the explicit motion description between two frames, so the temporal CNN will not need to estimate the motion implicitly. Figure 2 shows the process of extracting optical flow for two consecutive frames by using the Lucas-Kanade Algorithm [38].

However, we must consider contentious motion and not instantaneous, and to this purpose, several consecutive optical flow maps are concatenated to make a cuboid and fed to the CNN. The optical flow maps for every two frames could be calculated using the following optical flow constraint equation [39]:

$$I_x u + I_y v + I_t = 0 \tag{1}$$

where I_x, I_y, and I_t are the spatial-temporal image brightness derivatives, u is the horizontal optical flow, and v is the vertical optical flow [40]. To solve this problem,

[38] breaks down an image into sections denoted by Ω and then performs a weighted least-square fit of the Equation (1) to the model $[u\,v]^t$ shown in:

$$\sum_{x\in\Omega} W^2[I_x u + I_y v + I_t]^2 \tag{2}$$

where W is the window function which helps to minimise the below formula:

$$\begin{bmatrix} \sum W^2 I_x^2 & \sum W^2 I_x I_y \\ \sum W^2 I_y I_x & \sum W^2 I_y^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = - \begin{bmatrix} \sum W^2 I_x I_t \\ \sum W^2 I_y I_t \end{bmatrix}. \tag{3}$$

A filter $[1 \quad -1]$ is used to compute I_t between image 1 and image 2. First, using $\frac{[-1\ \ 8\ \ 0\ \ -8\ \ 1]}{12}$ kernel and its transpose form, I_x and I_y are computed. Second, using an isotropic element kernel, I_x, I_y, I_t will be smoothed down. Then, Equation (4) is solved for each pixel and eigenvalues of A are computed. Lastly, the eigenvalues (λ_i) are compared to the threshold (τ) constant that is determined at the beginning for noise reduction, and the results fall under the three conditions stated in (6), (7), (8). Under Condition 1, Cramer's rule [39] is used to solve the equations, for Condition 2 to compute u and v, the gradient flow has to be normalised, and for Condition 3, the optical flow u and v are zero. If we consider the video sequence as I_t where $1 \leq t \leq T$ with a subject figure on each frame of the sequence and T is the number of frames in the video.

$$IFA = \begin{bmatrix} a & b \\ b & c \end{bmatrix} = \begin{bmatrix} \sum W^2 I_x^2 & \sum W^2 I_x I_y \\ \sum W^2 I_y I_x & \sum W^2 I_y^2 \end{bmatrix}, \tag{4}$$

$$\lambda_i = \frac{a+c}{2} \pm \frac{\sqrt{4b^2 + (a-c)^2}}{2} \quad i = 1, 2, \tag{5}$$

$$\text{Condition 1}: \ \lambda_1 \geq \tau \text{ and } \lambda_2 \geq \tau, \tag{6}$$

$$\text{Condition 2}: \ \lambda_1 \geq \tau \text{ and } \lambda_2 < \tau, \tag{7}$$

$$\text{Condition 3}: \ \lambda_1 < \tau \text{ and } \lambda_2 < \tau. \tag{8}$$

After getting the output OF, the video sequence might have different duration in the temporal plane, and since CNN needs a fixed-sized input, we use the method shown in [9] to make a fixed size of 60 pixels × 60 pixels for the OF maps and feed this size to the cropping and resizing block to crop and resize the raw image frames before feeding the spatial-STC stream. We use 25 frames of the OF maps and the processed images for the input of our STC block since in most of the state-of-the-art datasets, including the ones used for accruing results in this paper, 25 frames cover a complete gait cycle [41, 42, 43]. To increase the training samples, we use Equation (9) to compute spatial displacement of ±5 in all directions and then

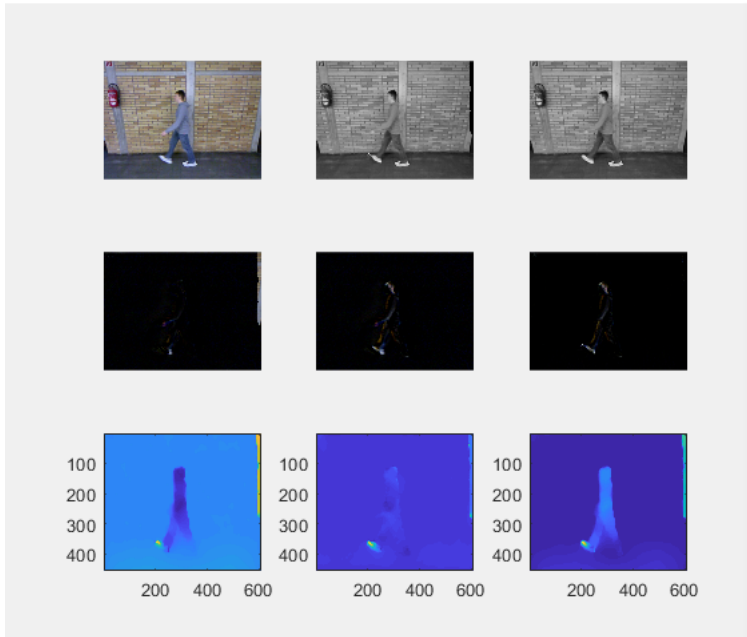

Figure 2. Optical Flow Maps for two consecutive frames

compute the corresponding mirror sequences on both processed raw images and OF maps:

$$FTD = \left(OF^{H\times W} \pm 5\right) - \left(\frac{\sum_{j=1} X_j}{N}\right). \tag{9}$$

If OF is an optical flow map with the height and width of H and W, respectively, and we have N images (X) in total, a single feedable training data before sending to the STC-Temporal stream is computed as FTD. A similar equation is used to pre-process raw images to feed our STC-Spatial stream. This operation will increase our data to around 540k samples with dimensions of $60 \times 60 \times 50$ before feeding into our STC block. Figure 3 shows the inner workings of our pre-processing block.

As shown in Figure 1, our Spatial-Temporal Convolution (STC) block will produce a combination of features, and then the results will be sent to an STA block to focus more spatial and temporal information. The STC layer outputs a feature vector that eliminates some interference and encodes some features of the person's appearance and movement, and the STA will focus our attention on the specific areas and frames in which our subjects are walking. The STA is only used to refine and improve the accuracy of the STC's output. By performing two fusions (explained in Section 3.2), STC learns the best temporal and spatial features simultaneously and then uses the attention layer to focus our attention to certain features by getting

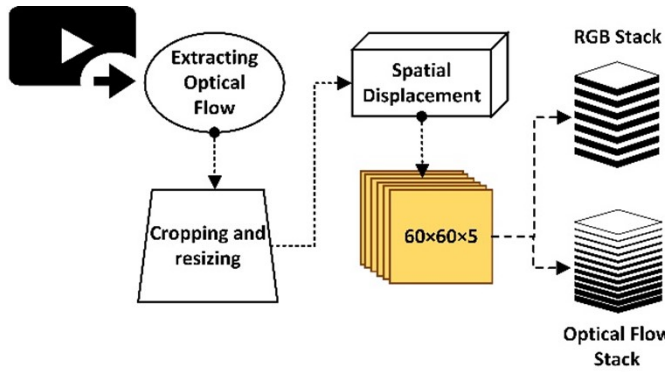

Figure 3. The proposed pre-processing block to prepare the datasets

the best regions automatically and assigning weights to each frame by employing long short-term memory (LSTM) (this will be explained in Section 3.3).

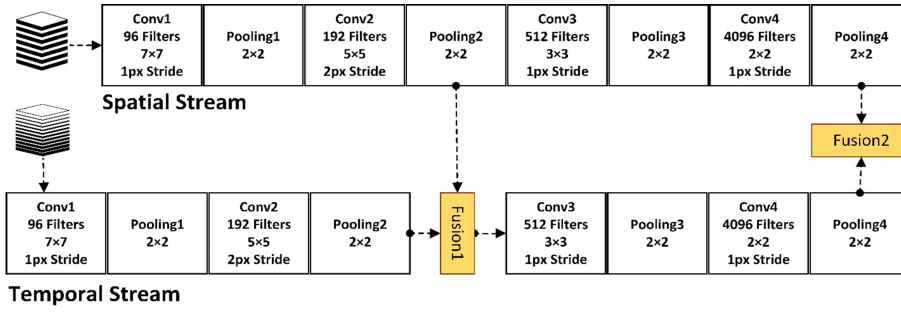

Figure 4. Our proposed two stream convolutional neural network – the STC block

3.1 STC Block

Video analysis owes its advances to the research into image analysis methods, including action recognition algorithms [44]. These methods typically take the local spatial-temporal features and perform shallow and high dimensional encoding like detecting interest points and pool them with the methods such as pyramid pooling introduced in [45]. Video can be divided into spatial and temporal components where the spatial will give information about appearances, actions and objects and temporal will give the information about movements either in the objects or the camera angle. To utilise both of these collections of information, we will divide these two jobs into two streams much like [34, 46] and use multi-layer fusion techniques to combine them. Where to fuse the networks is explained in Section 3.2.

Our two-stream network in the STC layer is illustrated in Figure 4. The input of the spatial stream consists of pre-processed RGB images (cropped and stacked video frames). The temporal stream is fed by pre-processed optical flow. Illustrated streams are entirely independent with two different inputs, but have the same structure, each of which has 4 convolutional layers accompanied by 4 max-pooling layers after each convolutional layer and a single normalisation after the first convolutional layer.

Since the size of our inputs for each stream are $60 \times 60 \times 50$ obtained from 25 raw RGB and another 25 OF frames the CNN was composed in a way that the first convolutional layer has 96 filters with size 7×7, with stride of 1 px, the second one has 192 filters with size 5×5 and 2 px stride, the third layer has 512 filters with size 3×3 and stride of 1 px and convolutional layer 4 contains 4 096 filters with size 2×2 and stride of 1 px. The pooling layers after each convolutional layer are 2×2. Both inputs of the networks have the same shape and dimensions as mentioned above, but the input of the spatial stream is a stack of RGB images, and the temporal stream is optical flow. Two fusion operations (explained in Section 3.2) make the information ready for the attention block.

3.2 Fusion Operation

This section introduces different spatial and temporal fusion techniques and our proposed architecture for this problem. There are several different methods to fuse two CNNs. At first glance, the problem of inaccurate spatial pixel-wise correspondence might not seem very important for gait human re-identification, but the differentiation between the motions such as walking, running, jumping from one frame to the next is very much dependent on action recognition. In other words, we want to fuse our two individual CNNs at a convenient layer that the spatial and temporal maps correspond on the same pixel. This problem can be solved by stacking the layers between the two networks. However, the corresponding channels are still an issue. Suppose that the special network channels are responsible for locating different body parts, including hands, legs and torso, and some channels are responsible for the periodic motion fields for each body part. Figure 5 shows an instance where temporal periodic motion was captured in correspondence with spatial by a webcam using fusion.

If f is a fusion function that fuses X_t^a and X_t^b which are two feature maps at time t then f has the output of $y_t \in R^{H'' \times W'' \times D''}$, from $X_t^a \in R^{H \times W \times D}$ and $X_t^b \in R^{H' \times W' \times D'}$ by performing a fusion operation $f : X_t^a, X_t^b \to y_t$ where H, W, and D, are Height, Width and the number of channels for corresponding feature maps, respectively. The fusion operation f applies to different layers of two CNNs in different ways, including multiple layers (multiple layer fusion), late and early fusion using simple operations put into symbolic $y = f(X_t^a, X_t^b)$.

To decide where to fuse the networks, we first go through a range of feasible fusion methods introduced in the literature using mathematical operations. We can first consider the sum operation, which will sum the feature maps at the same

Figure 5. Captured Periodic Motion using a low-quality webcam

location in space indicated by (i, j) and channel d. So $y = X^a_{i,j,d} + X^b_{i,j,d}$ where i, j and $d \geq 1$ and less than their respective height, width, and number of channels. However, this method needs considerably more training since the number of channels is chosen randomly. Hence, making the correspondence more accurate requires a lot of training and optimisation. Max fusion denoted in the equation below has a similar impact and consequently arbitrary correspondence.

In convolutional fusion, the feature maps will be stacked at first, and the output would be calculated through a convolution operation with $1 \times 1 \times 2D$ filters $fl \in R^{1\times1\times2D\times D}$ and bias of $b \in R^D$. The output naturally will be calculated as follows, where y^{cat} is the result of the concatenation operation explained above. In this case, fl can learn the correspondences of X^a_t and X^b_t and improve the performance and accuracy by minimising a loss function:

$$y = y^{cat} * fl + b. \tag{10}$$

In bi-linear fusion, for each of the (i, j) locations, a matrix is calculated using the outer products of the two feature vectors, and then all the locations are added together. The output is shown in Equation (11). In simple words, the output is the multiplicative interactions in corresponding (i, j) locations:

$$y = \sum_{i=1}^{H} \sum_{j=1}^{w} X^{a}_{i,j}{}' \otimes X^b_{i,j}. \tag{11}$$

As a result of this fusion, channels of both CNNs are combined, but on the other hand, we will lose all the spatial information at this point since they will be outputted as a bi-linear vector, as shown in [47]. The vector is fed into a SoftMax

layer for classification in this work. This method cannot be used in our approach since the spatial information will be lost, and we need this information for use in our STA block. In concatenation fusion, where the feature maps are stacked in the (i, j) for channel d as such that the output will be in 2 dimensions y_{2d} and y_{2d-1}, as shown below where X^a and X^b are two feature maps (with height, width and channel number of H, W, D, respectively) produced by the layers and belong to $R^{H\times W\times D}$ and the fused feature map is represented by $y \in R^{H\times W\times 2D}$ and $1 \leq i \leq H$, $1 \leq j \leq W$. This method will find the correspondence in the subsequent layers so that the network learns suitable filters for the job:

$$y_2 d = X^a_{i,j,d}. \tag{12}$$

Convolutional, max and sum fusion in the ReLU5 layers are employed as mentioned in [34]. Hence, we used the concatenation fusion on the second and fourth pooling layers to fuse our STC block's two spatial and temporal streams. To decide where to fuse the CNNs, we can use the concept of "upconvolutional" layers or zero-padding the smaller feature map [48, 49] since X^a_t and X^b_t have to have identical dimensions. The intent is to fuse the network at layers to keep the correspondence of temporal motion features and spatial features throughout the network. Which method of fusion to use, and in which layer the fusion happens, it can seriously impact the classification accuracy of the method. Later in Section 4, we will show different scenarios tested on different datasets, which will help us find the proper fusion technique and layer.

3.3 STA Block

In the real world, when we look at a human being from afar and try to recognise them, we usually focus on distinctive features in their movement as well as distinguishing their characteristics. In other words, we divert our attention to certain regions in a scene to find our saliency points [50]. There were several pieces of research on directing attention with saliency maps in the literature, including but not limited to [51, 52, 53]. Furthermore, spatial-temporal attention networks have recently been used for video and image analysis. [54] uses selective focus on RGB videos. In [55] a method was proposed that labelled every frame according to the actions performed in the scene and placed very dense labelling over the video frames. The model would get several frames as input and then assign weights to each frame. To our knowledge, only a few projects are working on person re-identification using attention models [56, 57] and no research that uses it for fine-tuning gait features.

In our approach, a recurrent Long-term, short-term (LSTM) was used much like [54] to produce temporal attention based on the velocity of motion (walking speed), a concept taken from [58] and produce spatial attention focusing on specific regions in the frame. We considered using a simple RNN since it is a popular model for feature extraction. Figure 6 shows a neuron used in RNN where h_t is the output

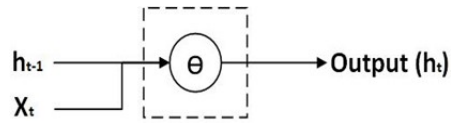

Figure 6. Structure of a typical RNN neuron

response at time step t, is based on the input X_t and the recursive outputs of the network itself denoted by h_{t-1} (hidden state), is calculated as:

$$h_t = \theta(W_{xh}^T X_t + W_{hh}^T h_{t-1} + b_h) \tag{13}$$

where θ is any nonlinear activation function, the connection vectors are W_{hh} and W_{xh}, and b_h represents the bias [56]. However, research has proven that with using RNN encoders alone, we might encounter the problem of vanishing gradient when modelling long-term temporal features of information sequences (lose a lot of temporal information). To rectify this issue, we use LSTM [59, 60] that is an advanced RNN architecture. A typical LSTM neuron structure is shown in [56, 61, 54, 62] where c_t is a memory cell which has edges of weight 1 which are self-connected and occur repeatedly. There are three gates denoted o_t, f_t, and i_t which are called output gate, forget gate and input gate, respectively, and can decide to read, reset or write to the c_t. Also g_t and h_t are used as an input modulation gate. The hidden state, h_t is calculated by element-wise multiplication of o_t and the calculated hyperbolic tangent of c_t. Hyperbolic tangent is denoted by $\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ in these calculations.

Other gates are calculated as follows, where W is a learnable connection vector and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid none-linear function between 0 and 1:

$$o_t = \sigma\left(W_{ox} X_t + W_{oh} h_{t-1} + b_o\right), \tag{14}$$

$$f_t = \sigma\left(W_{fx} X_t + W_{fh} h_{t-1} + b_f\right), \tag{15}$$

$$i_t = \sigma\left(W_{ix} X_t + W_{ih} h_{t-1} + b_i\right), \tag{16}$$

$$g_t = \phi\left(W_{gx} X_t + W_{gh} h_{t-1} + b_g\right). \tag{17}$$

In our method we introduce a Speed Gate, o_t, which will be calculated with attention to the velocity of movement (walking speed) of a subject in the leg area and applies it to the LSTM architecture. Our LSTM neuron is depicted in Figure 7 and formulated as follows:

$$vel_t = \sigma(W_{velx} X_t + W_{velh} h_{t-1} + b_{vel}). \tag{18}$$

Notice that vel_t has no authority to decide when to read, reset or write to the c_t and it is simply added and updated with the input gate. Now the memory cell and hidden state can be calculated from the below formulas using the gates above.

Where $\odot$ is the element-wise multiplication:

$$c_t = f_t \odot c_{t-1} \odot i_t \odot g_t + i_t \odot vel_t. \tag{19}$$

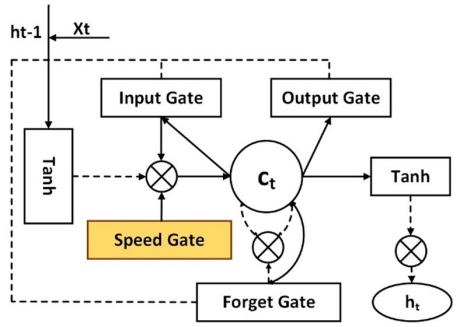

Figure 7. Neuron architecture for the LSTM used in the architecture

We take the output of the STC block, consider it as a feature map and insert it into the Attention block. This feature map has all the spatial and temporal information that we collected using the two-stream network from the input $X^{H\times W\times D}$. The recursive nature of the proposed model can focus the attention of the network to different regions in the person of interest. Now assume that $X^{H\times W\times 2D}$ is the input of our attention block after the STC and that the spatial part of the model only pays attention to the person of interest and ignores all the other spatial and temporal information. In order to do that, the LSTM should predict specific SoftMax locations and perform an operation and then perform a spatial pooling to combine the feature slices.

If the SoftMax locations have the size $H^s \times W^s$ the locations at time t are calculated by Υ_t as follows where w is the weight:

$$\Upsilon_{t,i} = \frac{e^{W_{t,i}\times h_{t-1}}}{\sum_{j=1}^{H^s\times W^s} e^{W_{t,j}\times h_{t-1}}}. \tag{20}$$

The spatial pooling operation naturally will be performed using Equation (21) where $X_{t,i}$ is the i^{th} feature slice at time t:

$$X_t^P = \sum_{i=1}^{H^s\times W^s} \Upsilon_{t,i} X_{t,i}. \tag{21}$$

For the temporal information to be taken into account we can take all the spatial information from the spatial pooling and import it into temporal pooling layer much

like [31] where v_s is vector characteristics in a video:

$$v = \frac{1}{T}\sum_{t=1}^{T} h_t. \tag{22}$$

In gait re-identification, we must focus on the global temporal structure instead of exploiting the local structure. Some papers exploit both [63]. Like them, instead of a simple averaging strategy, we will take the dynamic weighted some of the temporal feature vectors such that:

$$v = \sum_{t=1}^{T} \alpha_t h_t \tag{23}$$

where $\sum_{t=1}^{T} \Upsilon_t = 1$ and α_t is computed at each t inside of the LSTM encoder. The attention weight, Υ_t, at time step t, shows the relevance of the i^{th} temporal feature. Thus, we devised an equation that takes the input and gives us a relevance score e_t:

$$e_t = \sigma(W_X X_t + W_h h_{t-1} + b). \tag{24}$$

At the end we normalise e_t to obtain for all frames where $1 \leq t \leq T$ to obtain α_t:

$$\alpha_t = \frac{e^{e_t}}{\sum_{j=1}^{T} e^{e_j}}. \tag{25}$$

4 PERFORMANCE EVALUATION

To be able to test our architecture, we designed our experiment based on [31]. We then validate our approach based on the results obtained on four selected datasets suitable for gait re-identification. These datasets are introduced in this section, along with our experimental results presented against other states of the art. To present the results, first we have to introduce the four benchmark datasets we used for this experiment namely CASIA [64], and TUM Gait from Audio, Image and Depth (GAID) [42], PRID2011 [43], and iLIDS Video re-Identification (iLIDS-VID) [65, 1, 66, 67].

4.1 CASIA

The Institute of Automation Chinese Academy of Sciences has provided CASIA's *A*, *B* and *C* datasets. Dataset *A* contains 20 subjects with 12 image sequences (4 sequences for each of the three directions to the image plane). The length of each sequence is not identical for the variation of the walker's speed, but it must range from 37 to 127. The CASIA dataset includes 19 139 images and has size 2.2 GB [12]. Dataset B is a large multi-view gait database created in January 2005. There are 124 subjects, and the gait data was captured from 11 views. Three variations, namely view angle, clothing and carrying condition changes, are separately

considered. An infrared (thermal) camera collected dataset C in July and August 2005. It contains 153 subjects and considers four walking conditions (normal walking, slow walking, fast walking, and normal walking with a bag). The videos were all captured at night.

For testing, we chose to use CASIA Dataset B. The 124 subjects were videoed for 3 to 5 seconds in the same location and with the resolution of 320 px by 240 px from 0 degrees to 180 degrees with 18 steps. The video frame rate is around 25 fps, and subjects are either with no extra condition, carrying a bag/backpack or wearing a coat/jacket. Ten videos were captured from each view: six with normal conditions, two with a jacket/coat and two carrying a bag or backpack. It is important to mention that we are not using all the viewpoints included in this dataset. Only 90 degree angle videos were used with a split of 48.3 % (60) for the training and dev set and 51.7 % (64) for the test set. Other research with high accuracy does not split the subjects into training and testing sets. In [68] the split the videos for each person that gave them a high accuracy level within the dataset, but it does not apply to practical situations.

4.2 TUM-GAID

The TUM Gait from Audio, Image and Depth (GAID) dataset consists of 3 370 sequences of 305 individuals recorded with a Microsoft Kinect sensor in two sessions in a 3.5 m wide hallway. Therefore, three streams are available: depth, audio and video (colour). The resolution for audio and depth are 640×480 pixels with an approximately 30 fps frame rate. So, to use the dataset with the same method, we resized the frames to the exact resolution as the CASIA dataset, which is 320×240 pixels. The database was completed with each subject's gender, age, height and shoe type. A variety of walking conditions has been implemented for each subject in TUM-GAID, including six with normal conditions, two with covered shoes and two with a backpack. This dataset makes the feature extraction easy as it has been tested with the following algorithms with great results:

1. The Gait Energy Image (GEI),
2. GEI on Depth data (depth-GEI),
3. Gait Energy Volume (GEV) and
4. Depth Gradient Histogram Energy Image (DGHEI).

4.3 iLIDS-VID

UK Home Office has developed the Imagery Library for Intelligent Detection Systems (i-LIDS) from pedestrians passing an airport arrival hall using two cameras, which aims to stimulate the development of VA systems. Through the i-LIDS initiative, the Home Office assesses and promotes VA development for Event Detection scenarios (e.g. illegally parked vehicles) and Object Tracking scenarios (e.g. people

in airports) that are key to UK Government requirements. The i-LIDS datasets are widely regarded as the most comprehensive of their kind and have achieved substantial recognition since their launch in 2006. The iLIDS-VID has 600 images of 300 people and has one pair of camera views from the two cameras per individual. It is particularly challenging and needs more pre-processing time since various image frames are in each sequence. Also, there are similarities in clothes and lousy lighting to be considered.

4.4 Prid2011 Dataset

This dataset was created in the co-operation with the Austrian Institute of Technology to test person re-identification approaches. The dataset consists of images extracted from multiple person trajectories recorded from two static surveillance cameras. Images from these cameras contain a viewpoint change and a stark difference in illumination, background and camera characteristics. It has 749 people videoed with two camera views in an outdoor scene with few people.

4.5 Experiment Design

In order to train the network and predict the identity of the subjects, we needed to define a simple SoftMax function used in various works [34, 31]. If we have k number of subjects, x is the feature vector, and q is the person's identity, the function is as follows:

$$I(x) = P(q = c|x) = \frac{\exp(W_c x)}{\sum_k \exp(W_k x)}. \tag{26}$$

We define a W matrix for the SoftMax weights in which W_c and W_k are the c^{th} and k^{th} columns, respectively, for the loss function to be defined as follows:

$$L(x^a, x^b) = \frac{W_s Eu(V^+) + W_t Eu(V^-)}{W_t + W_s} + I(x^a) + I(x^b). \tag{27}$$

W_t are the temporal weights, and W_s are the spatial weights taken from our STC. This equation will take into account the Siamese loss based on Euclidian distance between two characteristic vectors $V = (x^a, x^b)$ [19, 69]. We treated the data through the pre-processing block for each dataset individually to train the model. The pre-processing block randomised the data and broke them into 80 %, 20 % proportion training and test sets, respectively. We used gradient descent to train the model with 1 000 epochs and a learning rate of 0.001.

We conduct our experiment based on the work of [31] which sends RGB images as the input of a single stream CNN and adds the optical flow using an RNN network. As mentioned before, we divided our architecture into several blocks in our experiment. We focus on the variations of the STC and STA and use a combination of these blocks with different methods to conduct the experiment. The experiment was run based on a single stream CNN or STC-1, a two-stream CNN (STC-2) with

Index	Method	iLIDS-VID				PRID2011				TUM-GAID		CASIA	
		R1	R5	R10	R20	R1	R5	R10	R20	R1	R20	R1	R20
M1	STC-1	57.8	82.6	83	89	70	82.8	94.9	93.4	80	95	35	62.2
M2	STC-1-STA	62.5	86.6	84.1	90.5	72	78.5	91.8	95.4	85	98	42	43.2
M3	STC-2-SF-4	53.2	90	88.6	70.1	52	88.2	91.2	90.8	82	97.3	48	44
M4	STC-2-SF-4-1	52	84.9	89.2	72.2	43	83.2	89.4	94.8	75	92.6	50	55
M5	STC-2-SF-4-2	43	91.4	89.5	69	53.2	87.1	94.7	95.4	73.2	95.2	52.3	50.3
M6	STC-2-SF-4-STA	45	88.8	90.1	77.4	54	87.2	90.5	95.6	77	98	68	52
M7	STC-2-SF-4-1-STA	68	90.9	92.8	75.2	68	83.6	88.9	95.1	75	95.3	54	86
M8	STC-2-SF-4-2-STA	65.8	91.4	80.4	70	65.8	84	93	93.6	76	86	64	73
M9	STC-2-CVF-4	64	90.8	80.9	82	64	88.5	92.3	95.1	68	88	56	75
M10	STC-2-CVF-4-1	44	89	81.2	85.2	56	84.5	93	96.9	69	96	45.5	70
M11	STC-2-CVF-4-2	45.5	87.2	79.9	87.6	45.5	81.8	94.7	98.5	83	86	48	86
M12	STC-2-CVF-4-STA	48	84	93.8	87.4	48	91.7	97.1	94.9	82.5	98	49.9	63
M13	STC-2-CVF-4-1-STA	36	83	94.2	89.4	49.9	83.2	94.6	90.8	49.9	84.1	35	63.5
M14	STC-2-CVF-4-2-STA	35	87.9	84	93	35	85.9	88.9	92.9	35	90	71.5	62.2
M15	STC-2-CF-4	51.5	90	92.1	90.1	71.5	94.9	91.9	79.9	71.5	86	68.6	83.2
M16	STC-2-CF-4-1	55.6	91.1	89.2	89.2	68.6	96.4	90.8	93.8	68.6	92	66	85.2
M17	STC-2-CF-4-2	57	92	94.5	89.9	66	83.6	97.6	96.6	66	97	69.9	80
M18	STC-2-CF-4-STA	62.9	89	94.2	97.6	69.9	91.5	90.4	94.4	69.9	86	68.2	86.8
M19	STC-2-CF-4-1-STA	70.2	90.6	92.8	95	71.2	93.5	91.6	96.6	95.5	98.9	71.2	87
M20	STC-2-CF-4-2-STA	70.4	92.8	96	97.8	87.3	97.1	98.2	99.7	99.2	99.5	78.3	88.3

Table 1. Methods used in our experiment – 20 variations of our architecture were used in the experiment

a single fusion of the fourth pooling layers or two fusions on pooling layers one and four or pooling layers two and four. Three different fusion techniques were also taken into account. Namely, sum fusion is denoted as SF, convolution fusion is denoted as CVF, and concatenation fusion is denoted as CF. These make twenty variations explained in Table 1. The results were presented on four datasets introduced above based on the ranking system in [31].

State of the Art Method	R1	R5	R10	R20
CNN + XQDA	53	81.4	–	95.1
RNN + OF	58	84	91	96
TSMRRNN	59.4	89.8	97.3	99.1
JSTRNN	55.2	86.5	–	97
STA-LSTM	64.8	90.7	96.4	98.3
Ours (M20)	**70.4**	**92.8**	**96**	**97.8**

Table 2. Comparison of our proposed approach with iLIDS-VID dataset

We keep the training length at 25-time steps when introducing the probe and gallery sequences. We just took the whole sequence for some cases where the gallery and probe length were more than the actual sequence. Probe sequences of length k are taken from the first k frames of the sequence recorded by the 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, and in some cases, probe and gallery sets were taken from the same camera. Table 1 shows Rank 1, Rank 5, Rank 10 and Rank 20 performances evaluation of

State of the Art Method	R1	R5	R10	R20
CNN + XQDA	77.3	93.5	–	99.3
RNN + OF	70	90	95	97
TSMRRNN	78.7	95.2	97.6	99.2
JSTRNN	79.4	94.4	–	99.2
STA-LSTM	78.3	96.7	99.3	99.7
Ours(M20)	**87.3**	**97.1**	**98.2**	**99.7**

Table 3. Comparison of our proposed approach with Prid2011 dataset

State of the Art Method	R1	R5	R10	R20
CNN + XQDA	90.4	–	–	92.1
RNN + OF	98.7	–	–	97.5
TSMRRNN	97.5	–	–	98.2
JSTRNN	98	–	–	97.1
STA-LSTM	99.3	–	–	99.2
Ours(M20)	**99.2**	**–**	**–**	**99.5**

Table 4. Comparison of our proposed approach with TUM-GAIT dataset

our 20 variations over the four datasets.

We used the Cumulative Mach Scoring (CMS) for evaluation, considering the classification problem M classes and N inputs. The output would be an $M \times N$ array of distances from N to M. Ordering and decoding the indexes of this gives the actual classification. An M by N matrix of class labels where the first column is the closest label and the M^{th} column is the furthest label. Taking the first column of this matrix, we will have a $1 \times N$ matrix. The cumulative match score is a correct result within the first m columns for each row. The cumulative match characteristic is the sum of these rows.

So, we use the correct quantity of their subjects and introduce it as Rank m. In this paper, several state-of-the-art methods for person re-identification were compared with our method. These results are provided we simulated almost every method and showed our experiment results in Table 1. Tables 2, 3, 4 and 5 show these results compared with the state of the art and correspond to Table 1.

State of the Art Method	R1	R5	R10	R20
CNN + XQDA	70	–	–	85
RNN + OF	72.2	–	–	85.2
TSMRRNN	65	–	–	79.9
JSTRNN	71.6	–	–	83.2
STA-LSTM	79.1	–	–	88
Ours(M20)	**78.3**	**–**	**–**	**88.3**

Table 5. Comparison of our proposed approach with CASIA dataset

5 CONCLUSIONS

In real-world scenarios, the subjects passing through an elaborate surveillance network cannot be expected to act predictably. We might not even get a complete gait cycle from a person of interest in most cases – as demonstrated in the paper. Clothing variation or carry bags considerably impact the system's performance in re-identification problems. Other abnormalities, including the camera angle, significantly aggravate the intra-class variation. Moreover, the similarity between gait appearances of different people extracted from low-level information introduces inter-class variations, resulting in similar gait signatures in more complex cases. Therefore, irregular gait recognition concerning viewpoint variations still needs the particular attention.

In this paper, the problem of video-based person re-identification was studied extensively. Most of the possible methods were simulated, and new and improved architecture with three blocks was introduced, using pre-processing block, a two-stream spatial-temporal CNN with two fusions, and a spatial-temporal LSTM attention network. We used the STC to extract the special and temporal features and the STA to make a weighted sequence of frames with specific weight assignments for each frame. A wide range of experiments has been conducted in this paper to achieve the best combination of these methods on four of the most popular public datasets. The outcome of our proposed end to end architecture was reflected against several state-of-the-art models with excellent results.

REFERENCES

[1] Ma, X.—Zhu, X.—Gong, S.—Xie, X.—Hu, J.—Lam, K. M.—Zhong, Y.: Person Re-Identification by Unsupervised Video Matching. Pattern Recognition, Vol. 65, 2017, pp. 197–210, doi: 10.1016/j.patcog.2016.11.018.

[2] Zhao, A.—Qi, L.—Li, J.—Dong, J.—Yu, H.: A Hybrid Spatio-Temporal Model for Detection and Severity Rating of Parkinson's Disease from Gait Data. Neurocomputing, Vol. 315, 2018, pp. 1–8, doi: 10.1016/j.neucom.2018.03.032.

[3] Chen, C.—Liang, J.—Zhao, H.—Hu, H.—Tian, J.: Frame Difference Energy Image for Gait Recognition with Incomplete Silhouettes. Pattern Recognition Letters, Vol. 30, 2009, No. 11, pp. 977–984, doi: 10.1016/j.patrec.2009.04.012.

[4] Boulgouris, N.—Hatzinakos, D.—Plataniotis, K.: Gait Recognition: A Challenging Signal Processing Technology for Biometric Identification. IEEE Signal Processing Magazine, Vol. 22, 2005, No. 6, pp. 78–90, doi: 10.1109/MSP.2005.1550191.

[5] BenAbdelkader, C.—Cutler, R.—Davis, L.: Stride and Cadence as a Biometric in Automatic Person Identification and Verification. Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, 2002, pp. 372–377, doi: 10.1109/AFGR.2002.1004182.

[6] Yoo, J. H.—Hwang, D.—Moon, K. Y.—Nixon, M. S.: Automated Human Recognition by Gait Using Neural Network. 2008 First Workshops on Image Processing Theory, Tools and Applications, 2008, pp. 1–6, doi: 10.1109/IPTA.2008.4743792.

[7] Nash, J. M.—Carter, J. N.—Nixon, M. S.: Dynamic Feature Extraction via the Velocity Hough Transform. Pattern Recognition Letters, Vol. 18, 1997, No. 10, pp. 1035–1047, doi: 10.1016/S0167-8655(97)00128-1.

[8] Chang, X.—Hospedales, T. M.—Xiang, T.: Multi-Level Factorisation Net for Person Re-Identification. 2018 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2109–2118, doi: 10.1109/CVPR.2018.00225.

[9] Castro, F. M.—Marín-Jiménez, M. J.—Guil, N.—Pérez de la Blanca, N.: Automatic Learning of Gait Signatures for People Identification. In: Rojas, I., Joya, G., Catala, A. (Eds.): Advances in Computational Intelligence (IWANN 2017). Springer, Cham, Lecture Notes in Computer Science, Vol. 10306, 2017, pp. 257–270, doi: 10.1007/978-3-319-59147-6_23.

[10] Han, J.—Bhanu, B.: Individual Recognition Using Gait Energy Image. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 28, 2006, No. 2, pp. 316–322, doi: 10.1109/TPAMI.2006.38.

[11] Xu, D.—Huang, Y.—Zeng, Z.—Xu, X.: Human Gait Recognition Using Patch Distribution Feature and Locality-Constrained Group Sparse Representation. IEEE Transactions on Image Processing, Vol. 21, 2012, No. 1, pp. 316–326, doi: 10.1109/TIP.2011.2160956.

[12] Hu, M.—Wang, Y.—Zhang, Z.—Zhang, D.—Little, J. J.: Incremental Learning for Video-Based Gait Recognition with LBP Flow. IEEE Transactions on Cybernetics, Vol. 43, 2013, No. 1, pp. 77–89, doi: 10.1109/TSMCB.2012.2199310.

[13] Yang, K.—Dou, Y.—Lv, S.—Zhang, F.—Lv, Q.: Relative Distance Features for Gait Recognition with Kinect. Journal of Visual Communication and Image Representation, Vol. 39, 2016, pp. 209–217, doi: 10.1016/j.jvcir.2016.05.020.

[14] Dikovski, B.—Madjarov, G.—Gjorgjevikj, D.: Evaluation of Different Feature Sets for Gait Recognition Using Skeletal Data from Kinect. 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, pp. 1304–1308, doi: 10.1109/MIPRO.2014.6859769.

[15] Farenzena, M.—Bazzani, L.—Perina, A.—Murino, V.—Cristani, M.: Person Re-Identification by Symmetry-Driven Accumulation of Local Features. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2360–2367, doi: 10.1109/CVPR.2010.5539926.

[16] Cheng, D. S.—Cristani, M.—Stoppa, M.—Bazzani, L.—Murino, V.: Custom Pictorial Structures for Re-Identification. Proceedings of the British Machine Vision Conference, BMVA Press, 2011, pp. 68.1–68.11, doi: 10.5244/C.25.68.

[17] Weinberger, K. Q.—Blitzer, J.—Saul, L.: Distance Metric Learning for Large Margin Nearest Neighbor Classification. In: Weiss, Y., Schölkopf, B., Platt, J. (Eds.): Advances in Neural Information Processing Systems 18 (NIPS 2005). MIT Press, Vol. 18, 2005, pp. 1473–1480, https://proceedings.neurips.cc/paper/2005/file/a7f592cef8b130a6967a90617db5681b-Paper.pdf.

[18] Hirzer, M.—Roth, P. M.—Köstinger, M.—Bischof, H.: Relaxed Pairwise Learned Metric for Person Re-Identification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.): Computer Vision – ECCV 2012. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7577, 2012, pp. 780–793, doi: 10.1007/978-3-642-33783-3_56.

[19] Hadsell, R.—Chopra, S.—LeCun, Y.: Dimensionality Reduction by Learning an Invariant Mapping. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, 2006, pp. 1735–1742, doi: 10.1109/CVPR.2006.100.

[20] Ding, S.—Lin, L.—Wang, G.—Chao, H.: Deep Feature Learning with Relative Distance Comparison for Person Re-Identification. Pattern Recognition, Vol. 48, 2015, No. 10, pp. 2993–3003, doi: 10.1016/j.patcog.2015.04.005.

[21] Yi, D.—Lei, Z.—Liao, S.—Li, S. Z.: Deep Metric Learning for Person Re-Identification. 2014 22nd International Conference on Pattern Recognition, 2014, pp. 34–39, doi: 10.1109/ICPR.2014.16.

[22] Roth, P. M.—Hirzer, M.—Köstinger, M.—Beleznai, C.—Bischof, H.: Mahalanobis Distance Learning for Person Re-Identification. In: Gong, S., Cristani, M., Yan, S., Loy, C. C. (Eds.): Person Re-Identification. Springer, London, Advances in Computer Vision and Pattern Recognition, 2014, pp. 247–267, doi: 10.1007/978-1-4471-6296-4_12.

[23] Ma, L.—Yang, X.—Tao, D.: Person Re-Identification over Camera Networks Using Multi-Task Distance Metric Learning. IEEE Transactions on Image Processing, Vol. 23, 2014, No. 8, pp. 3656–3670, doi: 10.1109/TIP.2014.2331755.

[24] Held, D.—Thrun, S.—Savarese, S.: Deep Learning for Single-View Instance Recognition. 2015, doi: 10.48550/arXiv.1507.08286.

[25] McLaughlin, N.—Del Rincon, J. M.—Miller, P.: Data-Augmentation for Reducing Dataset Bias in Person Re-Identification. 2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2015, pp. 1–6, doi: 10.1109/AVSS.2015.7301739.

[26] Rama Varior, R.—Wang, G.—Lu, J.—Liu, T.: Learning Invariant Color Features for Person Reidentification. IEEE Transactions on Image Processing, Vol. 25, 2016, No. 7, pp. 3395–3410, doi: 10.1109/TIP.2016.2531280.

[27] Li, W.—Zhao, R.—Xiao, T.—Wang, X.: DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159, doi: 10.1109/CVPR.2014.27.

[28] Pigou, L.—van den Oord, A.—Dieleman, S.—Van Herreweghe, M.—Dambre, J.: Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video. International Journal of Computer Vision, Vol. 126, 2018, No. 2, pp. 430–439, doi: 10.1007/s11263-016-0957-7.

[29] Ng, J. Y. H.—Hausknecht, M.—Vijayanarasimhan, S.—Vinyals, O.—Monga, R.—Toderici, G.: Beyond Short Snippets: Deep Networks for Video Classification. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4694–4702, doi: 10.1109/CVPR.2015.7299101.

[30] Jia, Y.—Huang, C.—Darrell, T.: Beyond Spatial Pyramids: Receptive Field

Learning for Pooled Image Features. 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3370–3377, doi: 10.1109/CVPR.2012.6248076.

[31] McLaughlin, N.—Martinez del Rincon, J.—Miller, P.: Recurrent Convolutional Network for Video-Based Person Re-Identification. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1325–1334, doi: 10.1109/CVPR.2016.148.

[32] Zhang, W.—Yu, X.—He, X.: Learning Bidirectional Temporal Cues for Video-Based Person Re-Identification. IEEE Transactions on Circuits and Systems for Video Technology, Vol. 28, 2018, No. 10, pp. 2768–2776, doi: 10.1109/TCSVT.2017.2718188.

[33] Yu, S.—Jia, S.—Xu, C.: Convolutional Neural Networks for Hyperspectral Image Classification. Neurocomputing, Vol. 219, 2017, pp. 88–98, doi: 10.1016/j.neucom.2016.09.010.

[34] Simonyan, K.—Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (Eds.): Advances in Neural Information Processing Systems 27 (NIPS 2014). Curran Associates, Inc., 2014, pp. 568–576, https://proceedings.neurips.cc/paper/2014/file/00ec53c4682d36f5c4359f4ae7bd7ba1-Paper.pdf.

[35] Sánchez, A.—Ruiz, J.—Moreno, A.—Montemayor, A.—Hernández, J.—Pantrigo, J.: Differential Optical Flow Applied to Automatic Facial Expression Recognition. Neurocomputing, Vol. 74, 2011, No. 8, pp. 1272–1282, doi: 10.1016/j.neucom.2010.07.017.

[36] Meng, Z.—Fu, S.—Yan, J.—Liang, H.—Zhou, A.—Zhu, S.—Ma, H.—Liu, J.—Yang, N.: Gait Recognition for Co-Existing Multiple People Using Millimeter Wave Sensing. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 849–856.

[37] Nickel, C.—Busch, C.—Rangarajan, S.—Möbius, M.: Using Hidden Markov Models for Accelerometer-Based Biometric Gait Recognition. 2011 IEEE 7th International Colloquium on Signal Processing and Its Applications, 2011, pp. 58–63, doi: 10.1109/CSPA.2011.5759842.

[38] Lucas, B. D.—Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI '81), Vol. 2, 1981, pp. 674–679.

[39] Barron, J. L.—Fleet, D. J.—Beauchemin, S. S.: Performance of Optical Flow Techniques. International Journal of Computer Vision, Vol. 12, 1994, No. 1, pp. 43–77, doi: 10.1007/BF01420984.

[40] Yedjour, H.: Optical Flow Based on Lucas-Kanade Method for Motion Estimation. In: Hatti, M. (Ed.): Artificial Intelligence and Renewables Towards an Energy Transition (ICAIRES 2020). Springer, Cham, Lecture Notes in Networks and Systems, Vol. 174, 2021, pp. 937–945, doi: 10.1007/978-3-030-63846-7_92.

[41] Barnich, O.—Van Droogenbroeck, M.: Frontal-View Gait Recognition by Intra- and Inter-Frame Rectangle Size Distribution. Pattern Recognition Letters, Vol. 30, 2009, No. 10, pp. 893–901, doi: 10.1016/j.patrec.2009.03.014.

[42] Hofmann, M.—Geiger, J.—Bachmann, S.—Schuller, B.—Rigoll, G.: The TUM Gait from Audio, Image and Depth (GAID) Database: Multimodal Recog-

nition of Subjects and Traits. Journal of Visual Communication and Image Representation, Vol. 25, 2014, No. 1, pp. 195–206, doi: 10.1016/j.jvcir.2013.02.006.

[43] Hirzer, M.—Beleznai, C.—Roth, P. M.—Bischof, H.: Person Re-Identification by Descriptive and Discriminative Classification. In: Heyden, A., Kahl, F. (Eds.): Image Analysis (SCIA 2011). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 6688, 2011, pp. 91–102, doi: 10.1007/978-3-642-21227-7_9.

[44] Laptev, I.—Marszalek, M.—Schmid, C.—Rozenfeld, B.: Learning Realistic Human Actions from Movies. 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8, doi: 10.1109/CVPR.2008.4587756.

[45] He, K.—Zhang, X.—Ren, S.—Sun, J.: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37, 2015, No. 9, pp. 1904–1916, doi: 10.1109/TPAMI.2015.2389824.

[46] Feichtenhofer, C.—Pinz, A.—Zisserman, A.: Convolutional Two-Stream Network Fusion for Video Action Recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1933–1941, doi: 10.1109/CVPR.2016.213.

[47] Lin, T. Y.—RoyChowdhury, A.—Maji, S.: Bilinear CNN Models for Fine-Grained Visual Recognition. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1449–1457, doi: 10.1109/ICCV.2015.170.

[48] Laina, I.—Rupprecht, C.—Belagiannis, V.—Tombari, F.—Navab, N.: Deeper Depth Prediction with Fully Convolutional Residual Networks. 2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 239–248, doi: 10.1109/3DV.2016.32.

[49] Dosovitskiy, A.—Brox, T.: Inverting Visual Representations with Convolutional Networks. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4829–4837, doi: 10.1109/CVPR.2016.522.

[50] Goferman, S.—Zelnik-Manor, L.—Tal, A.: Context-Aware Saliency Detection. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2376–2383, doi: 10.1109/CVPR.2010.5539929.

[51] Torres-Tramón, P.—Hromic, H.—Walsh, B.—Heravi, B. R.—Hayes, C.: Kanopy4Tweets: Entity Extraction and Linking for Twitter. Proceedings of the 6th Workshop on 'Making Sense of Microposts' co-located with the 25th International World Wide Web Conference (WWW 2016), CEUR Workshop Proceedings, Vol. 1691, 2016, pp. 64–66.

[52] Ba, J.—Mnih, V.—Kavukcuoglu, K.: Multiple Object Recognition with Visual Attention. 2014, doi: 10.48550/arXiv.1412.7755.

[53] Bazzani, L.—Larochelle, H.—Torresani, L.: Recurrent Mixture Density Network for Spatiotemporal Visual Attention. 2016, doi: 10.48550/arXiv.1603.08199.

[54] Sharma, S.—Kiros, R.—Salakhutdinov, R.: Action Recognition Using Visual Attention. 2015, doi: 10.48550/arXiv.1511.04119.

[55] Yeung, S.—Russakovsky, O.—Jin, N.—Andriluka, M.—Mori, G.—Fei-Fei, L.: Every Moment Counts: Dense Detailed Labeling of Actions in Complex

Videos. International Journal of Computer Vision, Vol. 126, 2018, No. 2, pp. 375–389, doi: 10.1007/s11263-017-1013-y.

[56] Song, S.—Lan, C.—Xing, J.—Zeng, W.—Liu, J.: An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2017), AAAI Press, Vol. 31, 2017, pp. 4263–4270, doi: 10.1609/aaai.v31i1.11212.

[57] Haque, A.—Alahi, A.—Li, F. F.: Recurrent Attention Models for Depth-Based Person Identification. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1229–1238, doi: 10.1109/CVPR.2016.138.

[58] Aminian, K.—Robert, P.—Jequier, E.—Schutz, Y.: Estimation of Speed and Incline of Walking Using Neural Network. IEEE Transactions on Instrumentation and Measurement, Vol. 44, 1995, No. 3, pp. 743–746, doi: 10.1109/19.387322.

[59] Graves, A.—Mohamed, A. R.—Hinton, G.: Speech Recognition with Deep Recurrent Neural Networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649, doi: 10.1109/ICASSP.2013.6638947.

[60] Hochreiter, S.—Schmidhuber, J.: Long Short-Term Memory. Neural Computation, Vol. 9, 1997, No. 8, pp. 1735–1780, doi: 10.1162/neco.1997.9.8.1735.

[61] Donahue, J.—Hendricks, L. A.—Guadarrama, S.—Rohrbach, M.—Venugopalan, S.—Saenko, K.—Darrell, T.: Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2625–2634, doi: 10.1109/CVPR.2015.7298878.

[62] Huang, Z.—Xu, W.—Yu, K.: Bidirectional LSTM-CRF Models for Sequence Tagging. 2015, doi: 10.48550/arXiv.1508.01991.

[63] Yao, L.—Torabi, A.—Cho, K.—Ballas, N.—Pal, C.—Larochelle, H.—Courville, A.: Describing Videos by Exploiting Temporal Structure. Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4507–4515, doi: 10.1109/ICCV.2015.512.

[64] Casia Gait Database. 2005, `http://www.cbsr.ia.ac.cn/english/Gait`.

[65] Li, M.—Zhu, X.—Gong, S.: Unsupervised Person Re-Identification by Deep Learning Tracklet Association. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11208, 2018, pp. 772–788, doi: 10.1007/978-3-030-01225-0_45.

[66] Wang, T.—Gong, S.—Zhu, X.—Wang, S.: Person Re-Identification by Discriminative Selection in Video Ranking. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 38, 2016, No. 12, pp. 2501–2514, doi: 10.1109/TPAMI.2016.2522418.

[67] Wang, T.—Gong, S.—Zhu, X.—Wang, S.: Person Re-Identification by Video Ranking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.): Computer Vision – ECCV 2014. Springer, Cham, Lecture Notes in Computer Science, Vol. 8692, 2014, pp. 688–703, doi: 10.1007/978-3-319-10593-2_45.

[68] Chen, J.—Liu, J.: Average Gait Differential Image Based Human Recognition. The Scientific World Journal, Vol. 2014, 2014, Art. No. 262398, doi:

10.1155/2014/262398.

[69] Chung, D.—Tahboub, K.—Delp, E. J.: A Two Stream Siamese Convolutional Neural Network for Person Re-Identification. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1992–2000, doi: 10.1109/ICCV.2017.218.

Babak Rahi recently obtained his Ph.D. from Brunel University London. He received his M.Sc. in computer communication networks in 2016 from the same university. He gained extensive experience in the industry for more than ten years. His research interests are computer vision, human re-identification using gait recognition, machine learning, and deep learning. Currently, he has been working on attention mechanisms and convolutional neural networks for human re-identification.

Man Qi is Senior Lecturer and the Research Lead in computing at the Canterbury Christ Church University, UK. Her main research interests are in the areas of intelligent systems and applications, data analytics and computer graphics. She is Fellow of the British Computer Society (FBCS) and Fellow of the Higher Education Academy (FHEA). She has published over 70 research papers and is on the editorial board of 5 international journals. She has been an external Ph.D. examiner for a number of universities in the UK and Australia. She has served as Chair and Program Committee member for over 50 international conferences and has been a long-term reviewer for many international journals.