IET Biometrics

Research Article

IET Journals
The Institution of Engineering and Technology

Pose-based deep gait recognition

ISSN 2047-4938
Received on 5th March 2018
Revised 23rd August 2018
Accepted on 25th September 2018
doi: 10.1049/iet-bmt.2018.5046
www.ietdl.org

Anna Sokolova[1] ✉, *Anton Konushin*[1,2]

[1]*National Research University Higher School of Economics, 20 Myasnitskaya str., Moscow 101000, Russia*
[2]*Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia*
✉ *E-mail: ale4kasokolova@gmail.com*

Abstract: Human gait or walking manner is a biometric feature that allows identification of a person when other biometric features such as the face or iris are not visible. In this study, the authors present a new pose-based convolutional neural network model for gait recognition. Unlike many methods that consider the full-height silhouette of a moving person, they consider the motion of points in the areas around human joints. To extract motion information, they estimate the optical flow between consecutive frames. They propose a deep convolutional model that computes pose-based gait descriptors. They compare different network architectures and aggregation methods and experimentally assess various body parts to determine which are the most important for gait recognition. In addition, they investigate the generalisation ability of the developed algorithms by transferring them between datasets. The results of these experiments show that their approach outperforms state-of-the-art methods.

1 Introduction

Gait recognition is a computer vision problem that comprises the identification of an individual in a video using the motion of their body as the only source of information. Unlike face recognition or re-identification problems, gait recognition does not rely solely on an individual's appearance, which tends to make the task more complex. Physiological studies show that each person has his/her own unique manner of walking, which is difficult to forge; thus, a person can be identified by the gait.

Gait recognition methods have several advantages that make them usable in many applied problems. First, unlike the face, iris, or fingerprints, gait representation can be recorded without a subject's cooperation. The second advantage is that motion can be captured and a person can be recognised even from a video with low resolution. Such features are very significant in video surveillance, and thus, gait recognition has become an important field of inquiry, particularly in the security field, the primary goals of which are the control of access to restricted areas and the detection of people who have previously been captured on camera (e.g. criminals).

Despite the uniqueness of gait, there are many factors that can affect gait representation and make the problem more complicated. Gait can appear to differ depending on the viewing angle or the clothing worn by the subject. Furthermore, wearing different shoes or carrying heavy bags will change the gait itself, and a recognition algorithm should be stable to such changes.

The problem of gait recognition is closely related to several computer vision problems. First, it is an identification problem similar to face recognition, with the difference that in gait recognition the motion of the body, but not the appearance, is of importance. A person may wear a mask veiling the face or a coat hiding the figure, but the body motion will still be the same. In addition, as well as action recognition it is a video classification problem, thus, the gait recognition problem may be solved by the same methods as action recognition (e.g. [1]). A third problem approximated to gait recognition is re-identification (re-id) [2]; however, as in face recognition, re-id usually addresses appearance rather than motion not requiring high frame rate and smoothness of the motion.

The similarity of these problems means that we can use approaches from adjacent fields in gait recognition. Most modern computer vision methods are based on convolutional neural networks (CNNs) and can be transformed for gait recognition. However, despite the development of such algorithms the most successful gait recognition approaches are still not deep and use handcrafted features. In this work, we propose a CNN-based algorithm for recognition of people based on their walking manner. This method proves to be more stable in transfer learning and achieves higher classification accuracy than previous gait recognition models comparable to state-of-the-art approaches.

Since we investigate the motion of the body, the person's appearance should not be taken into the account. Accordingly, we can consider the optical flow (OF) as the main source of information and avoid the use of raw images. Our experiments reveal that such an approach does obtain sufficient data and produces successful results.

Unlike [3, 4] we use not just motion of the whole body to recognise the human, but consider the hierarchy of body parts from full body to exact joints and observe the motion of the points in these shrinking areas. Such an approach is novel in gait recognition and distinguishes our method among the others.

Besides this, while using the proposed recognition model we do not need to know exact viewing angles the videos are shot at, which is an advantage of our method in the context of multiview gait recognition.

The main contributions of this study are as follows:

- We are the first to use part-based OF models considering the motion of the points in the surrounding of the human pose key points.
- Our approach achieves high recognition quality exceeding the state-of-the-art methods not only in side-view mode but in many settings of multi-view recognition.

2 Related work

Currently, there are two leading approaches to gait recognition. The more traditional approach is based on the extraction of handcrafted features from image frames. Most investigations following this approach use silhouette masks as the main source of information and extract features that show how these masks change. The most popular descriptor of gait used in such investigations is the gait energy image (GEI) [5], a binary mask averaged over the gait cycle of a human figure. This approach has developed significantly during recent years. Many different descriptors have been proposed

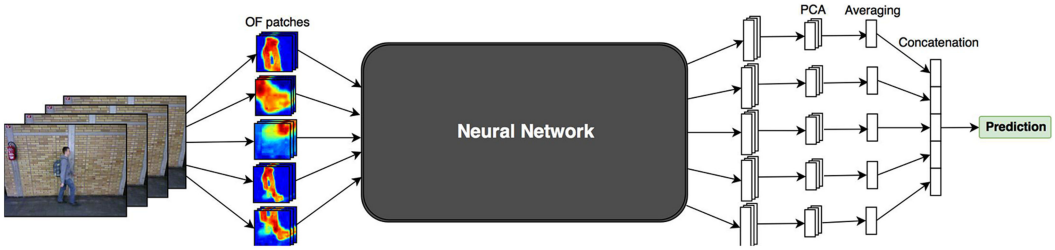

Fig. 1 *Pipeline of the algorithm*

for application to GEI (e.g. the histograms of oriented gradients [6, 7] and histograms of OF [8, 9]) or to entire silhouette sequences, e.g. the frame difference energy image [10] to provide additional aggregation for better gait representation. GEIs are also used in a unified metric learning framework [11] where joint intensity and spatial metric are optimised in order to mitigate the intrasubject differences and increase the intersubject ones.

One more gait representation technique based on silhouette extraction is entropy discrete fourier transform (EnDFT) [12]. This approach combines two other techniques: DFT [13] and gait entropy image (GEnI) [14]. In the first one, the mean of binary silhouette masks with exponentially decreasing weights is calculated, while the second one computes the entropy of each point of silhouette over the gait period. GEnI is the mixture of these approaches, the entropy is computed from DFT instead of GEI. Besides, the most effective human body parts are used for the recognition according to the recognition accuracy of the rows of DFT gait features.

Human body parts are considered in some other approaches, as well. Whytock *et al.* [15] propose a skeleton-based method considering the set of distance functions for a silhouette and in [16] the kinematic features based on the joints motion are fused with the spatial-temporal ones.

Despite the variety of hand-crafted features proposed for recognition, most of these approaches do not achieve high accuracy on the challenging gait databases. Nevertheless, there is a method that reaches the state-of-the-art result. The highest cross-view recognition quality is shown by Li *et al.* [17] who proposed the Bayesian approach. GEI is supposed to be a sum of gait identity values and some noise standing for different gait variations (view, clothing, carrying bags etc.) both following Gaussian distribution. Covariance matrices of these distributions are optimised by expectation-maximisation (EM) algorithm considering the joint distribution of GEI pairs. Similar to most of the multiview recognition approaches in [17] GEI images are calculated for each viewing angle separately, so, the knowledge of angles in the training set is required.

Another approach to gait recognition is neural networks. Deep models have been used to obtain the best results in most computer vision problems and have recently inspired new investigations of gait based on CNN. The similarity of gait and action recognition problems means that many approaches applied to the latter can be used for the former. The first and most classical deep method was proposed by Simonyan and Zisserman [18]. To recognise human actions, they trained a network with an architecture comprising two similar types of branches: image and flow streams. The former processes raw frames of video, while the latter gets the maps of OF computed from pairs of consecutive frames as input. To consider long-duration actions, several consecutive flow maps are stacked into blocks that are used as network inputs. Many action recognition algorithms based on this framework apply variant top architectures (including the recurrent architectures [19] and methods for fusing several streams [20]). Another method based on OF considers temporal information using three-dimensional convolutions [21].

The most applicable to gait recognition is the pose-based CNN [22] method proposed for action recognition. Similar to previous approaches, this method uses two streams but, in place of full-body maps, each stream receives patches in which different body parts are cropped. Thus, some joints are considered more precisely, which helps in the recording of small but important bodily motions.

The OF approach was applied to the gait recognition problem in several works [3, 4, 23], which proposed a deep model using blocks of OF maps containing full bodies as inputs to predict recorded individuals. Several other deep gait recognition solutions unite neural and GEI approaches. GEINet [24] computes gait energy images at various viewing angles for network input. In DeepGait [25] neural network features are extracted using silhouette masks as input and maximum response over the gait cycle is used. Wu *et al.* [26] proposed a deep CNN that predicted the similarity given a pair of gait sequences and considered different ways of comparing gait features. Siamese convolution network allowing to compare the pairs of objects is also used in one more state-of-the-art method, [27] where triplet and contrastive losses are considered for training, and in many other approaches. Being based on GEI computed for each view all these methods similarly to [17] need to know the angles of all the frames of the video. Such a requirement can hardly be accomplished in real life data.

Additionally to convolutional feed-forward neural networks the recurrent networks are applicable to gait recognition as well as to other problems concerning video. Tong *et al.* [28] use the long short-term memory unit (LSTM) [29] to get gait features from silhouette sequences, while several other approaches [30, 31] make intermediate pose evaluation and obtained pose features are fed into LSTM layers.

However, despite the success of neural network approaches such as these, some non-deep methods still achieve a higher quality of gait recognition.

The review of existing methods shows that in spite of the great variety of approaches, the problem remains unsolved. The main difficulty concerning multi-angle is not overcome and even side view recognition is still not perfect. Besides this, even achieving high quality on the existing benchmarks many methods require the knowledge of viewing angles of the videos, which complicates their applicability to real life data where the viewing angles are usually not labelled.

3 Proposed method

Here, we describe the pipeline of our proposed method. Although the algorithm is based on neural networks, two important data preprocessing stages are applied: motion map computation and frame-by-frame evaluation of the individual's pose. After these steps are completed, the network can be trained to classify video sequences. The scheme of our proposed approach is shown in Fig. 1.

Let us discuss all the stages of the algorithm.

3.1 Preprocessing the data

As our goal is to train a feature extractor that does not depend on, e.g. clothing colour or personal appearance, we can eliminate all colour information and use only motion. To do this, we compute maps of OF between each pair of consecutive frames and treat these maps as inputs. To calculate OF we use the method proposed by Farnebäck [32]. The algorithm consists of two steps: quadratic approximation of each pixel neighbourhood by polynomial expansion transform and point displacement estimation based on the transformation of the polynomials under translations. Having the vector field computed we consider three-channel OF maps in which the first two channels carry, respectively, horizontal and

vertical components of flow vectors and the third one carries magnitude. Prior to further processing, all maps are linearly transformed to the interval [0, 255] in a coding similar to that used in red, green and blue (RGB) channels. Since OF represents the apparent motion between frames, the only problem can appear if the human's clothing mimics the background, but as it actually happens very rarely we suppose that this problem does not limit our approach.

We can further assume that the motions of some parts of the human body are more informative than others and, correspondingly, evaluate the human pose and limit our analysis of OF maps to the neighbourhoods of such key positions. For example, the hands are not very informative, as while walking the human can move his hands independently (carrying a bag, speaking on the phone or just folding his arms). So, the features obtained from the hands' motion are supposed to be noisy and not useful for recognition. Thus, we expect that the bottom part of the body carries more gait information and therefore pay more attention to the legs than to either the hands or the head.

We, therefore, estimate the pose of the human in each frame by finding joint locations by OpenPose [33] algorithm. This method extends the convolutional pose machines [34] approach based on applying the cascade of several predictors specifying the estimations of each other. At each stage, there is a neural network with a small convolutional architecture predicting the heatmaps of joints locations. Such a step-by-step procedure allows us to increase the receptive field and observe the whole image as well as local features leading to prediction refinement.

Having found the locations of human pose key points, we crop five patches from the OF maps: right foot, left foot, upper body, lower body, and full body. Bounding boxes of these body parts are shown in Fig. 2. The idea of considering points motion maps in the neighbourhood of human body parts is novel compared to other gait recognition approaches based on OF.

The patches for network training are chosen in the following way. The leg patches are squares with key foot points in their centres, the patch should cover the whole foot including heel and toes, and thus, the side of the square is empirically chosen to be 25% of humans' height. The upper body patch contains all of the joints from the head to the hips (including the hands), and the lower body patch contains all of the joints from the hips to the feet (excluding the hands). Thus, each pair of consecutive frames produces five patches for use as network inputs. Prior to inputting the data into the network, the resolution of each patch is decreased to 48×48 pixels.

It is also worth noting that as we detect the person and his body parts in each frame, we filter the frames and take into account only those frames that contain the full human figure. So, sometimes several frames at the beginning and the end of the video (where the person is not in the frame entirely) are not used.

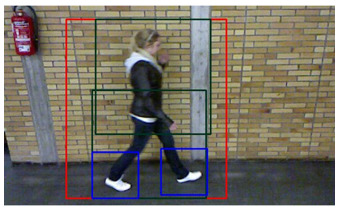

Fig. 2 *Bounding boxes for human body parts: right and left feet (small blue boxes at the bottom), upper and lower body (green boxes in the middle) and full body (big red box)*

3.2 Data augmentation

The primary component of the proposed algorithm involves the extraction of neural features. As every deep neural network has a large number of parameters, it is necessary to augment the data to obtain a stable and not over-fitted algorithm.

The data is augmented using classic spatial changes. In the training process, four numbers are uniformly sampled for each of the five considered body parts to construct the bounds of an input patch. The first two numbers are left and right extensions chosen from the interval $[0, w/3]$, where w is the width of the initial bounding box, while the latter two numbers are upper and lower extensions in $[0, h/3]$, where h is the height of the box. The resulting bounded patches are cropped from the OF maps and then resized to 48×48.

This augmentation allows us to obtain patches containing body parts undergoing both spatial shifts and zoom. If the sum of the sampled numbers is close to zero, a large image of the body part with a very little excess background is obtained; otherwise, the image is smaller with more background. On the other hand, fixing the sums of the first and the second pairs of bounds produces body parts with the same size but changed location inside the patch.

3.3 Training the neural network

The network is trained using the augmented data produced above and then used as a feature extractor, with the outputs of the last hidden layer used as gait descriptors. In the testing stage, instead of sampling random bounds for each patch, their mean value is taken ($w/6$ and $h/6$, respectively) to locate the body part in the centre of the patch.

3.3.1 CNN architectures and training methods: We consider and compare two network architectures. The first one is based on the VGG-19 [35] network but has one less convolutional block; details of this architecture are shown in Table 1.

Each column of this table corresponds to a block of layers. The first four blocks are convolutions, with each row denoting, for each layer in the block, the size of its filters (3×3 for all the layers) and the number of filters; the number of filters per block is doubled in each succeeding block. Additionally, there are four max-pooling layers of size 2×2 after each convolutional block.

The next two blocks are fully connected, with each comprising one linear dense layer of size 4096 and a dropout (d/o) with the probability parameter $p = 0.5$. Similar to convolutions, the dense layers are followed by rectified linear unit (ReLU) nonlinearities. The final column denotes the top block, comprising a dense layer and a softmax nonlinearity. The number of units in this block is equal to the number of training subjects that can train the network for a classification task.

During training, an L_2 norm of dense layer weights is added to the loss function to enhance the regularisation against overfitting.

Networks similar to this have produced the best results in previous experiments [3] in which the blocks of several consecutive OF maps were used as network inputs. However, that approach [3] did not take into account the key points of the human pose; instead, the full body was used. We will further compare this previous study with our approach in Section 4.

Similar to the block-based model [3], we trained our network step by step. We started from 1024 units in two hidden dense layers and trained the network while doubling the layer size each time the accuracy ceased to increase. When the layer size reached 4096 units, the training stopped. Each widening of the network was

Table 1 VGG-like architecture

B1	B2	B3	B4	F5	F6	SM
3 × 3, 64	3 × 3, 128	3 × 3, 256	3 × 3, 512	—	—	—
3 × 3, 64	3 × 3, 128	3 × 3, 256	3 × 3, 512	4096	4096	soft-
		3 × 3, 256	3 × 3, 512	d/o	d/o	max
pool 2	pool 2	pool 2	pool 2	—	—	—

Table 2 WideResNet architecture

B1	B2	B3	B4
3 × 3, 16 BN	$\begin{bmatrix} \text{BN} \\ 3\times3, 64 \\ \text{BN} \\ 3\times3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{BN} \\ 3\times3, 128 \\ \text{BN} \\ 3\times3, 128 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{BN} \\ 3\times3, 256 \\ \text{BN} \\ 3\times3, 256 \end{bmatrix} \times 3$
stride 1	stride 1	stride 2	stride 2

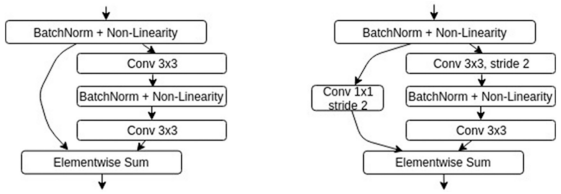

Fig. 3 *Details of the construction of residual blocks. Left: common block with the same input and output tensor shape. Right: 'decreasing block' with stride and extra convolution used when the number of filters increases and map size decreases*

implemented by random initialisation of new parameters to add extra regularisation and prevent overfitting while training.

After developing this deep CNN with both convolutional and dense layers, we considered implementing the second architecture not used in our previous research: a fully convolutional residual network. One of the more successful architectures in computer vision used for image classification task is the ResNet architecture, in which residual connectivity allows for the transfer of information from low to high levels and the addition of each new block increases the accuracy of classification. Furthermore, the absence of fully connected layers in the ResNet architectures reduces the number of parameters of the model. These features have made ResNet a very popular architecture and inspired our investigation of its use here. Although residual networks achieve great results, they require large numbers of layers to significantly improve their performance, and each new block significantly lengthens the training process. Another problem with deep networks is exploding or vanishing gradients as a result of very long paths from the last to the first layer during backpropagation. To avoid these problems, we use the wide residual network (WideReSNet) [36] with decreased depth and increased residual block width; the resulting reduction in the number of layers made the training process much faster and allowed the network to be optimised with less regularisation.

Table 2 shows the details of our WideResNet architecture.

Each column in the table defines a set of convolution blocks with the same number of filters. As in VGG-like architectures, all of the convolutions have kernels of size 3×3. The first layer has 16 filters and is followed by batch normalisation (BN) and then three residual blocks, each comprising two convolutional layers with 64 filters with normalisation between them. This block construction is classical for WideResNet architectures, and all of our blocks have similar structures. In each successive group, the blocks have twice as many filters as in the previous group. Each first convolutional layer in groups B3 and B4 has a stride with parameter 2; therefore, the tensor size begins at 48×48 pixels and decreases to 24×24 and then 12×12 in groups B3 and B4, respectively.

The detailed construction of residual blocks is shown in Fig. 3. The scheme on the left corresponds to a common block without strides when the input and output shapes coincide; that on the right defines the structure of the first block of the group (columns B3 and B4) as the number of filters doubles and the size of the map decreases. In this case, we add one auxiliary convolutional layer with a 1×1 kernel, stride, and doubled number of filters to make all of the shapes equal before the summation.

Following the residual blocks is a final process necessary to make the network useful for classification. This comprises one additional BN layer, an average pooling that 'flattens' the sequence of maps 12×12 obtained following the convolutions to one vector and one dense layer with softmax on the top. The number of units in this dense layer is equal to the number of subjects in the training set. All of the activations except for the final softmax are ReLUs that follow the BN layers.

3.4 Final classification

The network is trained to predict one of the subjects from the training set with a patch cropped from its OF map. As our goal is to construct a feature extractor that can be used without retraining or fine-tuning for any testing set, we use the outputs of the final hidden layer as the gait descriptors and construct a new classifier for them. As we assume that the gait descriptors of a given person are spatially close to each other, we can use one of the simplest methods, the nearest neighbour (NN) classifier. We make a further L_2 normalisation of all gait feature vectors prior to fitting and classifying, as many studies have shown that having a uniform length across all vectors improves the accuracy of NN classification.

Although most classical measures of distance between two vectors are Euclidean, we also consider the use of the Manhattan distance as a metric. So, we make experiments with each of these two metrics as similarity measure in the NN classifier and compare them. Despite the fact that normalisation is always performed relative to the L_2 distance, in most experiments it has been shown that finding the closest descriptor with respect to L_1 metrics produces better and more stable results.

To improve and speed up the classification, we reduce the dimensionality of the feature vectors using a principal component analysis (PCA) algorithm to reduce noise in the data and accelerate the fitting of classifiers.

3.4.1 Fusion of feature vectors: The trained neural network and fitted NN classifier allow us to predict which subjects have patches with the OF around one of the body parts. As we are starting the analysis from an initial video sequence, if we consider j body parts we obtain j patches for each pair of consecutive frames and therefore $(N-1)j$ descriptors for the video, where N is the number of frames in the sequence. If we consider the frames separately, we can obtain $(N-1)j$ answers per video; however, we require only one. We investigate two methods for constructing one feature vector from all network outputs. The first, a 'naive' method, involves averaging all of the descriptors by calculating the mean feature vector over all frames and all body parts. This approach is naive in that the descriptors corresponding to different body parts have different natures; even if we compute them using one network with the same weights, it would be expected that averaging the vectors would mix the components into a disordered result. Surprisingly, the accuracy achieved using this approach is very high, as will be shown through comparison with another approach in the next section.

The second and more self-consistent approach is averaging the descriptors over only time. After doing so, j mean descriptors corresponding to each of j body parts are obtained and concatenated to produce one final feature vector.

4 Data and experiments

4.1 Datasets

We evaluated the methods described in the preceding section on three popular gait databases: the 'TUM Gait from Audio, Image and Depth' (TUM-GAID) [37] dataset; CASIA Gait Dataset B [38]; and the OU-ISIR Gait Database, Large Population Dataset (OULP) [39].

The TUM-GAID dataset, comprising videos for 305 subjects, is a sufficiently large database for gait recognition problems. The videos have lengths of 2–3 s apiece at a frame rate of 30 fps; we used all of them in our experiments. All videos are recordings of people of full height walking, captured from the side view. Although each person is recorded from only one viewpoint, there are several video sequences per subject taken under various conditions, e.g. wearing different shoes or carrying various items.

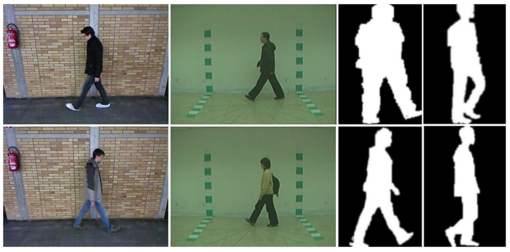

Fig. 4 *Examples of frames from three databases: TUM-GAID (the first column), CASIA Gait Dataset B (the second column) and OU-ISIR (the third and fourth columns)*

Table 3 Results on TUM-GAID dataset

Method	Evaluation	
Architecture, aggregation and metrics	Rank-1, %	Rank-5, %
VGG (PCA 1000), avg, L_2	96.4	100.0
VGG (PCA 1000), avg, L_1	97.8	100.0
VGG (PCA 500), concat, L_2	97.4	99.9
VGG (PCA 500), concat, L_1	98.8	**100.0**
wide ResNet (PCA 230), avg, L_2	98.3	99.9
wide ResNet (PCA 230), avg, L_1	99.2	99.9
wide ResNet (PCA 150), concat, L_2	98.8	99.8
wide ResNet (PCA 150), concat, L_1	**99.8**	99.9
VGG + blocks [3], L_1	97.5	99.9
CNN + support vector machine (SVM) [4], L_2	98.0	99.6
deep multi-task model (DMT) [23]	98.9	—
random subspace method (RSM) [43]	92.0	—
skeleton variance image (SVIM) [15]	84.7	—
divergence-Curl-Shear descriptor (DCS) [42]	99.2	—
H2M [42]	99.2	—
PFM [41]	99.2	99.5

Overall, there are ten videos per subject: six normal walks, two walks in coating shoes, and two walks with a backpack. Examples of video frames are shown in Fig. 4 (first column). As is true of most gait databases, TUM-GAID's records contain one person per video without any intersection between figures. This is, of course, a naive approach, as in real life people often walk together with their bodies intersecting; nevertheless, this structure allows for the training of a model for checking if the problem of gait recognition can be theoretically solved. As the TUM-GAID is a relatively large database, it was the main data source in our experiments.

The second database we investigated is CASIA Gait Dataset B. This dataset contains only 124 subjects but is multiview: records are captured from 11 different viewpoints at angles ranging from 0° to 180°. As in the TUM-GAID dataset, there are ten videos per person captured under different conditions: normal walking, carrying a bag, and wearing a coat. Despite the fact that there are several sequences per subject and viewpoint, the dataset is very small for the purposes of deep modelling, as neural networks must contain many parameters, especially if they are to handle multiview modes. Thus, in most of the experiments with CASIA dataset, we used only side-view videos captured from a constant angle and solved the side-view problem alone. Frames from the CASIA database are shown in the second column of Fig. 4. Although the combination of a great variety of conditions (a large number of viewpoints, carrying and clothing conditions) and relatively small number of subjects makes this database really challenging we have conducted an additional experiment concerning all viewing angles presented in this database to check if the proposed method can be generalised for multiview recognition.

The third dataset we used is the OU-ISIR Gait Database. This is the largest gait database of the three, containing over 4000 subjects, each captured by two cameras at four different angles (55°, 65°, 75°, and 85°). The dataset is formatted as a set of silhouette sequences, making the data different from those in the other sets. Examples of silhouettes from this database are shown in the third and fourth columns of Fig. 4. The specialty of the OU-ISIR dataset is that the shooting angles change smoothly from 55° to 85°. Thus, the frames shot at the intermediate angles are labelled both nearest 'main' angles. We applied our algorithm to the OU-ISIR database to determine if silhouette masks are sufficient for gait recognition. As the algorithm we use for pose estimation requires full images, we did not create body part patches and extracted only full body features. We used a subset of the OU-ISIR database comprising two walks taken from 1912 subjects to meet the protocols of benchmarks [40].

4.2 Performance evaluation

The experiments were conducted in the following manner. The feature extractor was trained on a training set containing all data for approximately one-half of all subjects in the database (155 for TUM-GAID, 64 for CASIA, and 956 for OU-ISIR). The rest of the subjects were used for fitting the final classifier and testing the overall algorithm. The fitting components comprised four normal walks per person, while the testing components contained a further six walks (including two pairs of walks under different additional conditions). We randomly sampled 64 training subjects from the CASIA base, while the splits for the TUM-GAID and OU-ISIR datasets were provided by their authors.

For each of the testing videos, the algorithm returns the vector of a probability distribution over all subjects from the testing set. We evaluated the quality of classification computing *Rank-1* and the *Rank-5* metrics that defined the ratio of videos in which the correct label was among the top five classifier answers. We also plotted cumulative match characteristics (CMC) curve to compare different techniques more clearly.

Additionally to identification experiments we compared the verification quality of different methods. All the videos for testing subjects were divided into training and testing parts the same way as in the identification task. For each pair of training and testing videos, the algorithm returns the distance between the corresponding feature vectors and predicts if there is the same person on both video sequences according to this distance. To evaluate the quality of the verification we plotted the receiver operating characteristic (ROC) curves of false acceptance rates and false rejection rates and calculated equal error rates (EERs).

4.3 Experiments and results

The goal of all experiments was to explore the influence of different conditions on gait performance, including:

- network architecture and aggregation methods;
- joints used for training and testing the algorithms;
- length of the captured walk.

In addition, we investigated the generality of the algorithms by training them on one dataset and applying them to another, as we expected that algorithms that depend only on body motion would work equally well on different databases.

We compared our results with those produced during the last few years including the approach based on Fisher vectors [41] and multimodal features (RGB frames, audio, and depth) [42], which have demonstrated state-of-the-art results.

The first set of experiments aimed at evaluating the approach itself and comparing different technical methods such as network architectures, aggregation methods, and similarity measures. All the algorithms were trained from scratch. The results are shown in Table 3, from which it is seen that our approach was most accurate and that it outperformed all the state-of-the-art methods.

In Table 3, 'Avg' denotes a naive approach for feature aggregation in which the mean descriptor is computed over all body parts, while 'concat' defines the concatenation of descriptors, which produces better results.

Table 4 Results for different conditions on TUM-GAID dataset

Method	Normal	Backpack	Shoes	Avg
WideResNet, L_1	**100.0**	**100.0**	**99.4**	**99.8**
VGG + blocks [3]	99.7	96.5	96.5	97.5
CNN + SVM L_2 [4]	99.7	97.1	97.1	98.0
DCS [42]	99.7	99.0	99.0	99.2
H2M [42]	99.4	100.0	98.1	99.2

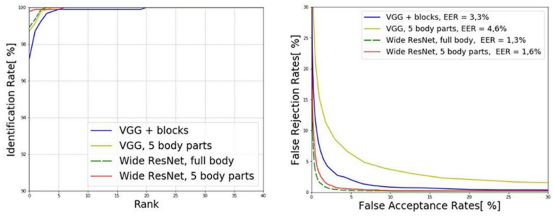

Fig. 5 *CMC (left) and ROC (right) curves for TUM-GAID dataset under different settings*

Table 5 Results on lateral-view part of CASIA dataset

Architecture, aggregation, and metrics	Rank-1, %
WideResNet (PCA 150), avg, L_2	85.1
WideResNet (PCA 130), avg, L_1	86.7
WideResNet (PCA 170), concat, L_2	84.8
WideResNet (PCA 170), concat, L_1	**93.0**
VGG + blocks [3], L_1	74.9

It is worth noting that in the pyramidal fisher motion (PFM) [41] approach, the input frames had initial size 640×480 and the resolution was not changed. Even though we reduced the inputs, the quality of the algorithms was not only maintained but even improved. Although the Rank-5 metric produced using the VGG-like network was larger, the difference was very small and the WideResNet architecture required far fewer parameters, suggesting that the latter architecture is more appropriate for solving the gait recognition problem.

It is also interesting to note that the L_1 metric always produced the higher accuracy of classification, suggesting that it is more suitable for measuring the similarity of gait feature vectors.

To check how sensitive the proposed method is to small appearance changes, we evaluated the recognition quality for different clothing and carrying conditions separately. The results of such separation are presented in Table 4. They confirm that although the changes in the silhouette and shoe change can influence not only appearance but the gait itself, our model can cope with them and remain fairly accurate.

To compare the identification and verification qualities of different settings graphically we plotted CMC and ROC curves and calculated EER. These metrics are shown in Fig. 5.

As the WideResNet architecture was dominantly successful in the evaluations, all further experiments were conducted using this network structure.

Table 5 shows a comparison of the performance on the lateral-view part of the CASIA dataset to the results using blocks of OF maps [3]. The accuracy is quite high despite the use of the rather poor training set; this might be attributable to the relatively small number of parameters in the WideResNet architecture and the corresponding lack of overfitting that could appear in the previous model [3].

Similar to Table 4, Table 6 shows what impact different clothing and carrying conditions presented in CASIA database have on recognition quality. The data variability in this collection is more complex than in TUM-GAID, as coat changes the silhouette more than coating shoes, thus the decomposition of results on this dataset into the components with different conditions is more informative. It is seen from the table that while the presence of the bag does not make recognition worse, the coat actually does. This worsening is

Table 6 Results for different conditions on CASIA dataset

Method	Normal	Bag	Coat	Avg
WideResNet, L1	**100.0**	**100.0**	**78.8**	**93.0**
VGG + blocks [3], L_1	94.5	78.6	51.6	74.9

Table 7 Comparison of average recognition rates for three angles on the CASIA database

Method	Average rank-1, %			
Model	Probe view			
	54	90	126	Mean
WideResNet (PCA 230), concat, L_1	**77.8**	**68.8**	74.7	**73.8**
SPAE [44]	63.3	62.1	66.3	63.9
Wu [26]	**77.8**	64.9	**76.1**	72.9

predictable as the coat is not just another shirt or blouse of a different style. Unlike shirts and blouses, a coat is outerwear which hides human figure, changes its shape, and makes the motion of many points of the body invisible. Thus such dress-up is a really challenging factor complicating gait recognition. Indeed, the shape change and joints hiding distinguish clothes difference from view variability which affects only inner gait representation of the model, but the gait appearance itself.

Nevertheless, the accuracy does not decrease dramatically remaining higher than the average accuracy of the method [3].

To check if our algorithm can be applied to multiview data we have conducted one more experiment with CASIA Dataset B. Following the protocol from [26] we have trained the network on all the available data for the first 24 subjects (ten videos per subject shot at all 11 angles) and then tested the feature extractor on the rest 100 subjects. Table 7 shows the average recognition rates at three probe angles: 54°, 90°, and 126°. The gallery angles are all the rest ten angles (excluding the corresponding probe one). We compare our results with [26] showing the state-of-the-art result on CASIA and stacked progressive auto-encoders (SPAE) [44] where one uniform model is trained for gait data with different viewing and carrying condition variations. For two of three considered angles, our method achieves the highest accuracy and even outperforms the other methods on the average.

In the experiments on the OU-ISIR database, we trained the network using all the maps and all the view angles for each training subject. During testing, we looked at gallery and probe views: as the gallery views, we used those fixed at 85°, while the probe views comprised all other angles. The NN classifier was fitted onto the gallery frames from the first walk and tested on the probe frames from the second walk.

We compare our method with several deep and non-deep approaches: view transformation models quality-dependent view transformation model (wQVTM) [45] and transformation consistency measures (TCM+) [46] methods linear discriminant analysis (LDA) [47] and multi-view discriminant analysis (MvDA) [40] based on discriminant analysis, Bayesian approaches [17, 48], and GEINet [24]. The results and a comparison are shown in Table 8.

The experimental results revealed that the proposed algorithm can be generalised to multiview and can obtain high accuracy even in cases of partial information. They show that even the absence of any texture inside the human figure (when the transitions between body parts and pieces of clothing are not visible) keeps recognition possible.

The second row in Table 8 shows the results of the network trained on all the available subjects except 956 testing ones. Such extended training set containing 2888 subjects allowed us to improve the quality and outperform the state-of-the-art method for some view angles. Despite being very high the results achieved using this extended set are not highlighted as they are obtained following the other training protocol.

Fig. 6 shows the ROC and CMC curves under the fixed gallery view 85°. The curves of other methods were provided by their authors. Although the recognition accuracy of our method is a bit

Table 8 Comparison of Rank-1 and EERs on OU-ISIR dataset obtained from silhouette masks

Method	Rank-1, %			EER, %		
Model	Probe view					
	55	65	75	55	65	75
WideResNet, L_1	92.1	96.5	97.8	**0.9**	**0.8**	**0.8**
WideResNet, L_2, extended training set	95.9	97.8	98.5	0.6	0.5	0.5
GEINet [24]	81.4	91.2	94.6	2.4	1.6	1.2
LDA [47]	87.5	96.2	97.5	6.1	3.1	2.3
MvDA [40]	88.0	96.0	97.0	6.1	4.6	4.0
wQVTM [45]	51.1	68.5	79.0	6.5	4.9	3.7
TCM + [46]	53.7	73.0	79.4	5.5	4.4	3.7
DeepGait + Joint Bayesian [48]	89.3	96.4	98.3	1.6	0.9	0.9
joint Bayesian [17]	**94.9**	**97.6**	**98.6**	2.2	1.6	1.3

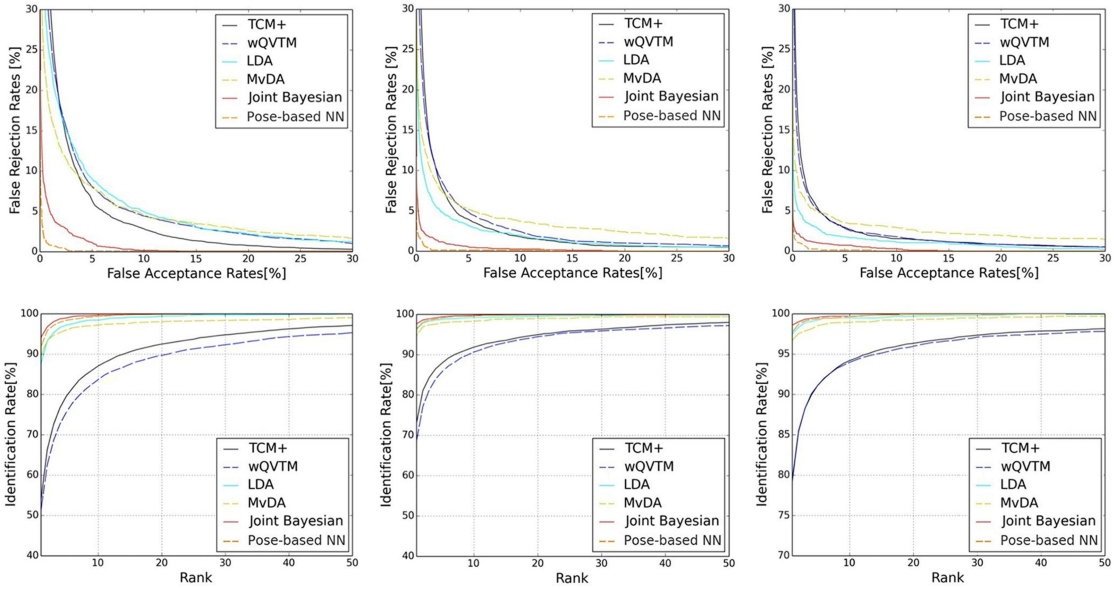

Fig. 6 *ROC (the first row) and CMC (the second row) curves under* 85° *gallery view and* 55°, 65°, *and* 75° *probe view, respectively*

Table 9 Comparison of Rank-1 on OU-ISIR dataset following 5-fold cv protocol

Method	Rank-1, %				
Model	Angular diff.				
	0	10	20	30	Mean
WideResNet, L_1	98.4	98.2	97.1	94.1	97.0
Takemura [27]	99.2	99.2	98.6	97.0	98.8
Wu [26]	98.9	95.5	92.4	85.3	94.3

Table 10 Comparison of the results on TUM-GAID dataset obtained using different parts of the body

Body parts	Rank-1, %	Rank-5, %
legs	79.7	86.5
upper body	96.2	99.7
lower body	96.3	99.6
full body	98.9	**100.0**
full body, upper body, lower body	99.4	**100.0**
full body, upper body, lower body, legs	**99.8**	99.9

lower than the best one [17], the quality of verification turns out to be higher.

To compare our approach with two more cross-view recognition methods [26, 27] that seem very successful on the OU-ISIR database, we trained the same model following the protocol from [26]. We conducted five-fold cross-validation (cv) using all 3844 subjects shot by two cameras. Similar to [27], we aggregated the results and considered various differences between probe and gallery views. The comparison with the two approaches mentioned above is presented in Table 9.

The results obtained after cv confirm the ability of the proposed method to recognise people at different viewing angles. The accuracy of the algorithm is a bit lower than [27] but outperforms [26]. Besides, as we mentioned above, unlike these methods we do not need the information that the frames have been shot at the same or different viewing angles.

After the experiments using different training techniques, we investigated which body parts are most important in gait recognition. Results using the OU-ISIR database revealed that full-body features are quite informative; therefore, we compared the models trained on distinct body parts. We trained the network in several additional modes: on each of body parts separately (full body, pairs of leg patches, upper, and lower body) and on three ‘ big’ and ‘ middle’ patches: full body, upper body, and lower body in order to check whether such a big scale is enough or particular legs consideration is really needed. The results for TUM-GAID are shown in Table 10.

It is seen from the first four rows that the larger the considered area is the higher accuracy is obtained. However, even the model

Table 11 Comparison of results on TUM-GAID for different lengths of videos

Length of video	Rank-1, %	Rank-5, %
50 frames	94.3	94.9
60 frames	97.5	97.8
70 frames	99.4	99.5
full length	99.8	99.9

Table 12 Comparison of results on CASIA-B for different lengths of videos

Video length	Multi-view protocol average rank-1, %				Side-view rank-1, %
	Probe view				
	54	90	126	Mean	90
50 frames	66.85	60.95	64.05	63.95	90.33
70 frames	67.05	62.70	66.50	65.42	91.38
90 frames	67.10	66.95	69.95	68.00	92.42
110 frames	75.25	68.45	73.20	72.30	92.68
full length	77.80	68.80	74.70	73.77	92.95

Table 13 Quality of transfer learning

Training set	Testing set	
	CASIA	TUM
CASIA	92.7	66.5
TUM	76.8	99.8

trained on full body patches is not perfect: classification error is more than 1%. The composition of three ‘ big’ and ‘ middle’ patches turns out to be more successful than the models based on distinct body parts, but its accuracy is still lower than the five-parts-based one. Thus we can conclude that not only one big patch, but the set of big and middle patches is not enough for good recognition and precise consideration of legs surroundings improves the classification.

Our third avenue of interest was determining the length of video needed for good gait recognition. In all of the preceding experiments, we used the entire video sequence, totalling up to 90 (for TUM-GAID) and 130 (for CASIA-B) frames per sequence. Tables 11 and 12 list the results of testing the algorithms on shortened sections of TUM-GAID and CASIA sequences (following both multi-view and side-view protocols described above). As the viewing angle changes smoothly in the OU-ISIR gait database and the duration of each period of constant view is about 30 frames, which is already quite short, this experiment cannot be done on this dataset. Since we train the network on distinct maps and do not aggregate the features obtained for each video while training, we have used the same WideResNet model trained on full-length videos in each of the experiments, but only the first n frames of each test video were used in the testing phase. We consider $n = 50, 60, 70$ for the TUM-GAID dataset and $n = 50, 70, 90, 110$ for the CASIA-B database as all the sequences there are longer. Since we use only those frames where the figure is fully visible and crop the boxes containing only the figure or special part of the human's figure, the starting and finishing positions are not important and do not have an influence on the recognition.

Although the length of the gait cycle is about 1 s, or 30–35 frames, such short sequences are found to be insufficient for good individual recognition. The experiments conducted on two datasets verify that increasing the number of consecutive frames for classification improves the results. The expanded time of analysis is required because body point motions are similar but not identical for each step and, correspondingly, using long sequences makes recognition more stable to small inter-step changes in walking style. It is worth noting that video sequences may consist of a non-integer number of gait cycles but still be accurately recognised.

In a final set of experiments, we examined the stability and transferability of the proposed algorithm. If the feature extractor is actually general and does not depend on the background or the appearance of a person, it should be able to extract features from videos even if they are recorded under conditions different from the initial video. To check such generality, we used both the available datasets in our experiment by attempting to train the algorithm on one of the databases and evaluating its quality on the other without fine-tuning. One database was used to train the neural network and to get feature extractor and the final classifier was fitted and tested on the second one. Since the OU-ISIR database is distributed in a silhouette form, the models trained on TUM and CASIA are not applicable to this dataset and we cannot transfer the models from the OU-ISIR to these two datasets and back. Thus, only two of three databases were used in this set of experiments. Table 13 shows the accuracy of the transference of the algorithm between datasets.

Note that even when the algorithm was trained on the other dataset, it worked on CASIA dataset B better than the method [3] trained on CASIA. Nevertheless, the accuracy of recognition deteriorated considerably following transference between databases, with the error increasing by a factor of three. With TUM-GAID, the results are even worse. Training the classifier on CASIA resulted in recognition of only 66,5% of the TUM testing videos, which is a factor of 1, 5 worse than that obtained by the algorithm trained on TUM. This suggests that the algorithm over fits and that the amount of training data (particularly in the CASIA dataset) is not sufficient for constructing a general algorithm. To make algorithm stable to small variations of camera parameters, its height or distance from the subject, we need more varied training dataset. All the databases existing now consist of videos shot in exactly the same conditions which prevent the generalisation of the algorithms.

5 Implementation details

Some of the auxiliary methods were implemented using public libraries. The bounding boxes for human figures were computed using silhouette masks found by background subtraction. This is quite a rough method, but as each frame in the databases contained only one moving person, it worked well. The maps of OF were calculated by applying the Farnebäck [32] algorithm implemented in the OpenCV library. The poses were evaluated using the open-source implementation of the [33] method to find the key points of the body. For the main part of the algorithm, we used Lasagne with a Theano backend and trained the networks on an NVIDIA GTX 1070 GPU. The main WideResNet employing five body parts was trained on the TUM-GAID dataset in 10 h. The model was optimised using the Nesterov Momentum gradient descent method with the learning rate reduced from 0.1 by a factor of 10 each time the training quality stopped increasing.

6 Conclusions and further work

In this study, we proposed a pose-based convolutional neural model for gait recognition. Our experimental results demonstrated

that although sufficiently high accuracy can be obtained only by using OF maps for the full-height region, collecting additional information from regions around the joints improves the results, surpassing the state-of-the-art on TUM-GAID. Our model can also be successfully applied to the moving silhouettes in OU-ISIR, which shows that the most important information for gait recognition is the movement of external edges and that our method can be straightforwardly applied to multiview gait recognition.

7 Acknowledgment

This work was supported by grant RFBR #16-29-09612 'Research and development of person identification methods based on gait, gestures, and body build in video surveillance data'.

8 References

[1] Karpathy, A., Toderici, G., Shetty, S., *et al.*: 'Large-scale video classification with convolutional neural networks'. Proc. 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'14), Columbus, Ohio, 2014, pp. 1725–1732

[2] Wang, X.: 'Intelligent multi-camera video surveillance: a review', *Pattern Recogn. Lett.*, 2013, **34**, (1), pp. 3–19

[3] Sokolova, A., Konushin, A.: 'Gait recognition based on convolutional neural networks'. Int. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Moscow, Russia, 2017, vol. XLII-2/W4, pp. 207–212

[4] Castro, F.M., Marín-Jiménez, M.J., Guil, N., *et al.*: 'Automatic learning of gait signatures for people identification', in Rojas, I., Joya, G., Catala, A. (Eds): '*Advances in computational intelligence*' (Springer International Publishing, Cham, 2017), pp. 257–270

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

[6] Dalal, N., Triggs, B.: 'Histograms of oriented gradients for human detection'. Proc. 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'05) – Volume 1, San Diego, CA, USA, 2005, pp. 886–893

[7] Liu, Y., Zhang, J., Wang, C., *et al.*: 'Multiple HOG templates for gait recognition'. Proc. 21st Int. Conf. on Pattern Recognition (ICPR2012), Tsukuba, JAPAN, 2012, pp. 2930–2933

[8] Laptev, I., Marszalek, M., Schmid, C., *et al.*: 'Learning realistic human actions from movies'. CVPR 2008 – IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, 2008, pp. 1–8

[9] Yang, Y., Tu, D., Li, G.: 'Gait recognition using flow histogram energy image'. 2014 22nd Int. Conf. on Pattern Recognition, Stockholm, Sweden, 2014, pp. 444–449

[10] Chen, J., Liu, J.: 'Average gait differential image based human recognition', *Sci. World J.*, 2014, **2014**, pp. 1–8

[11] Makihara, Y., Suzuki, A., Muramatsu, D., *et al.*: 'Joint intensity and spatial metric learning for robust gait recognition'. 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, 2017, pp. 6786–6796

[12] Rokanujjaman, M., Islam, M., Hossain, M.A., *et al.*: 'Effective part-based gait identification using frequency-domain gait entropy features', *Multimedia Tools Appl.*, 2015, **74**, pp. 3099–3120

[13] Makihara, Y., Sagawa, R., Mukaigawa, Y., *et al.*: 'Gait recognition using a view transformation model in the frequency domain'. Computer Vision – ECCV 2006, Graz, Austria, 2006, pp. 151–163

[14] Bashir, K., Xiang, T., Gong, S.: 'Gait recognition using gait entropy image'. Proc. 3rd Int. Conf. on Crime Detection and Prevention, London, UK, 2009, pp. 1–6

[15] Whytock, T., Belyaev, A., Robertson, N.M.: 'Dynamic distance-based shape features for gait recognition', *J. Math. Imaging Vis.*, 2014, **50**, (3), pp. 314–326

[16] Deng, M., Wang, C., Cheng, F., *et al.*: 'Fusion of spatial-temporal and kinematic features for gait recognition with deterministic learning', *Pattern Recognit.*, 2017, **67**, pp. 186–200

[17] Li, C., Sun, S., Chen, X., *et al.*: 'Cross-view gait recognition using joint Bayesian'. Proc. SPIE 10420, Ninth Int. Conf. on Digital Image Processing (ICDIP 2017), Hong Kong, China, 2017

[18] Simonyan, K., Zisserman, A.: 'Two-stream convolutional networks for action recognition in videos'. Proc. 27th Int. Conf. on Neural Information Processing Systems (NIPS'14), Montréal, CANADA, 2014, vol. 1, pp. 568–576

[19] Ng, J.Y.H., Hausknecht, M.J., Vijayanarasimhan, S., *et al.*: 'Beyond short snippets: deep networks for video classification'. 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, Massachusetts, 2015, pp. 4694–4702

[20] Feichtenhofer, C., Pinz, A., Zisserman, A.: 'Convolutional two-stream network fusion for video action recognition'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, 2016, pp. 1933–1941

[21] Varol, G., Laptev, I., Schmid, C.: 'Long-term temporal convolutions for action recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, **40**, pp. 1510–1517

[22] Chéron, G., Laptev, I., Schmid, C.: 'P-CNN: pose-based CNN features for action recognition'. Int. Conf. on Computer Vision, Santiago, Chile, 2015

[23] Marín-Jiménez, M., Castro, F., Guil, N., *et al.*: 'Deep multi-task learning for gait-based biometrics'. IEEE Int. Conf. on Image Processing (ICIP), Beijing, China, 2017

[24] Shiraga, K., Makihara, Y., Muramatsu, D., *et al.*: 'GEINet: view-invariant gait recognition using a convolutional neural network'. 2016 Int. Conf. on Biometrics (ICB), Halmstad, Sweden, 2016, pp. 1–8

[25] Zhang, X., Sun, S., Li, C., *et al.*: 'Deepgait: a learning deep convolutional representation for gait recognition'. Biometric Recognition, Shenzhen, China, 2017, pp. 447–456

[26] Wu, Z., Huang, Y., Wang, L., *et al.*: 'A comprehensive study on cross-view gait based human identification with deep cnns', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **39**, pp. 209–226

[27] Takemura, N., Makihara, Y., Muramatsu, D.: 'On input/output architectures for convolutional neural network-based cross-view gait recognition'. *IEEE Trans. Circuits Syst. Video Technol.*, 2017, **28**, pp. 1–1

[28] Tong, S., Fu, Y., Ling, H., *et al.*: 'Gait identification by joint spatial-temporal feature'. Biometric Recognition, Shenzhen, China, 2017, pp. 457–465

[29] Hochreiter, S., Schmidhuber, J.: 'Long short-term memory', *Neural Comput.*, 1997, **9**, (8), pp. 1735–1780

[30] Liu, D., Ye, M., Li, X., *et al.*: 'Memory-based gait recognition'. Proc. British Machine Vision Conf. (BMVC), York, UK, 2016, pp. 82.1–82.12

[31] Feng, Y., Li, Y., Luo, J.: 'Learning effective gait features using LSTM'. Int. Conf. on Pattern Recognition, Cancún, México, 2016, pp. 325–330

[32] Farnebäck, G.: 'Two-frame motion estimation based on polynomial expansion'. Image Analysis, Halmstad, Sweden, 2003, pp. 363–370

[33] Cao, Z., Simon, T., Wei, S.E., *et al.*: 'Real-time multi-person 2d pose estimation using part affinity fields'. Computer Vision and Pattern Recognition, Honolulu, Hawaii, 2017

[34] Wei, S.E., Ramakrishna, V., Kanade, T., *et al.*: 'Convolutional pose machines'. Computer Vision and Pattern Recognition, Las Vegas, Nevada, 2016

[35] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition'. CoRR, 2014, abs/1409.1556

[36] Zagoruyko, S., Komodakis, N.: 'Wide residual networks'. Proc. British Machine Vision Conf. (BMVC), York, UK, 2016, pp. 87.1–87.12

[37] Hofmann, M., Geiger, J., Bachmann, S., *et al.*: 'The TUM gait from audio, image and depth (GAID) database: multimodal recognition of subjects and traits', *J. Vis. Commun. Image Represent.*, 2014, **25**, (1), pp. 195–206

[38] Yu, S., Tan, D., Tan, T.: 'A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition'. Proc. 18th Int. Conf. on Pattern Recognition (ICPR), Hong Kong, China, 2006, vol. 4, pp. 441–444

[39] Iwama, H., Okumura, M., Makihara, Y., *et al.*: 'The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition', *IEEE Trans. Inf. Forensics Sec.*, 2012, **7**, (5), pp. 1511–1521

[40] Mansur, A., Makihara, Y., Muramatsu, D., *et al.*: 'Cross-view gait recognition using view-dependent discriminative analysis'. IJCB 2014 – 2014 IEEE/IAPR Int. Joint Conf. on Biometrics, Clearwater, FL, USA, 2014

[41] Castro, F.M., Marín-Jiménez, M., Medina-Carnicer, R.: 'Pyramidal fisher motion for multiview gait recognition'. 22nd Int. Conf. on Pattern Recognition, Stockholm, Sweden, 2014, pp. 1692–1697

[42] Castro, F.M., Marn-Jiménez, M.J., Guil, N.: 'Multimodal features fusion for gait, gender and shoes recognition', *Mach. Vis. Appl.*, 2016, **27**, (8), pp. 1213–1228

[43] Guan, Y., Li, C.T.: 'A robust speed-invariant gait recognition system for walker and runner identification', 2013, pp. 1–8

[44] Yu, S., Chen, H., Wang, Q., *et al.*: 'Invariant feature extraction for gait recognition using only one uniform model', *Neurocomputing*, 2017, **239**, pp. 81–93

[45] Muramatsu, D., Makihara, Y., Yagi, Y.: 'View transformation model incorporating quality measures for cross-view gait recognition', *IEEE Trans. Cybern.*, 2015, **46**, pp. 1602–1615

[46] Muramatsu, D., Makihara, Y., Yagi, Y.: 'Cross-view gait recognition by fusion of multiple transformation consistency measures', *IET Biometrics*, 2015, **4**, pp. 62–73

[47] Belhumeur, P.N., Hespanha, J.A.P., Kriegman, D.J.: 'Eigenfaces vs. Fisherfaces: recognition using class specific linear projection', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, (7), pp. 711–720

[48] Li, C., Min, X., Sun, S., *et al.*: 'Deepgait: a learning deep convolutional representation for view-invariant gait recognition using joint Bayesian', *Appl. Sci.*, 2017, **7**, p. 210