Neuroinformatics (2018) 16:457–471
https://doi.org/10.1007/s12021-018-9362-4

ORIGINAL ARTICLE

Learning Efficient Spatial-Temporal Gait Features with Deep Learning for Human Identification

Wu Liu[1] · **Cheng Zhang**[2] · **Huadong Ma**[1] · **Shuangqun Li**[1]

Published online: 6 February 2018

Abstract

The integration of the latest breakthroughs in bioinformatics technology from one side and artificial intelligence from another side, enables remarkable advances in the fields of intelligent security guard computational biology, healthcare, and so on. Among them, biometrics based automatic human identification is one of the most fundamental and significant research topic. Human gait, which is a biometric features with the unique capability, has gained significant attentions as the remarkable characteristics of remote accessed, robust and security in the biometrics based human identification. However, the existed methods cannot well handle the indistinctive inter-class differences and large intra-class variations of human gait in real-world situation. In this paper, we have developed an efficient spatial-temporal gait features with deep learning for human identification. First of all, we proposed a gait energy image (GEI) based Siamese neural network to automatically extract robust and discriminative spatial gait features for human identification. Furthermore, we exploit the deep 3-dimensional convolutional networks to learn the human gait convolutional 3D (C3D) as the temporal gait features. Finally, the GEI and C3D gait features are embedded into the null space by the Null Foley-Sammon Transform (NFST). In the new space, the spatial-temporal features are sufficiently combined with distance metric learning to drive the similarity metric to be small for pairs of gait from the same person, and large for pairs from different persons. Consequently, the experiments on the world's largest gait database show our framework impressively outperforms state-of-the-art methods.

Keywords Gait recognition · Siamese neural network · Spatio-temporal features · Metric learning · Human identification

Introduction

Biometrics based automatic human identification is one of the most fundamental and significant research topic in bioinformatics and computer vision field. Among the massive biometric authentication traits, the discrimination of human gait is strongly supported in the research of biomechanics, physical medicine studies, and psychological studies (Wang et al. 2003). As shown in Fig. 1, gait is the walking posture of a person, comprising a regular movement trend and variations present at joints of the upper limbs and lower limbs during walking. Gait recognition is a recognition technology based on biometrics (Ma and Liu 2017), and is intended to identify a person according to the walking posture of the person. With the gait analysis, we can accurately detect many biomedical information, such as gender, age, race and the like, of a person to whom the gait belongs can be obtained. Moreover, the gait information can also be utilized to diagnose and cure many diseases, such as Neurodegenerative Diseases (Ren et al. 2017; Xia et al. 2015), Parkinson's disease (Samà et al. 2017), Huntington's Disease (Mannini et al. 2016), and so on. For the field of recognition technology, gait is biometrics of great potential, which mainly manifests in the following three aspects: 1) *remotely accessible*: surveillant can obtain gait information of a specific subject from a distance, and collect it secretly in a contactless manner. Differently, the biometrics such

✉ Wu Liu
liuwu@bupt.edu.cn

Cheng Zhang
zhang.7804@osu.edu

Huadong Ma
mhd@bupt.edu.cn

Shuangqun Li
shuangqunli@bupt.edu.cn

[1] Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China

[2] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA

as iris and fingerprint are collected with the need of a person's cooperation. The remote access is very important in intelligent video surveillance. 2) *Robustness*: even in low resolution videos, a gait feature still works well. In contrast, an accurate face recognition (Chen et al. 2014; Chen et al. 2017) and vocal print recognition impose relatively high requirements on the quality of data resources. 3) *Security*: it is difficult to imitate or camouflage human gait. If a person changes his/her gait in public deliberately, he/she would become more suspicious and gain attention.

However, accurate gait recognition is still a challenging work as 1) the unconspicuous inter-class differences from different people; and 2) the large intra-class variations from the same person as the different walking speeds, viewpoints, clothing, and belongings. The detail challenges can be summarized as: 1) complexity surveillance environments — the human gait identification is very sensitive to cluttered environments, illumination changes, partial occlusions, and crowded people; 2) diverse subject-related factors — the different walking speed, dressing, and carrying conditions. all seriously influence the human gait and automatic identification ; 3) the cross-view variance — the cross-view variance leads to the appearances of human gait be substantially altered, which will increase the intra-class variations and decrease the inter-class variations.

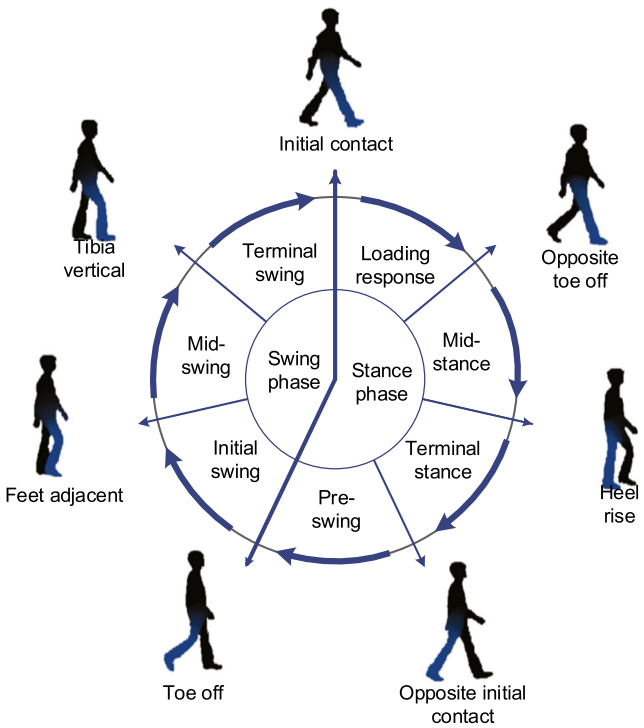

Fig. 1 A complete cycle of the human gait. Besides humen identification, the gait analysis can also be utilized to diagnose and cure many diseases, such as Neurodegenerative Diseases (Ren et al. 2017; Xia et al. 2015), Parkinson's disease (Samà et al. 2017), Huntington's Disease(Mannini et al. 2016), and so on

To solve these challenges, two kinds of gait recognition methods are studied: model-based and appearance-based. Model-based approaches (Ariyanto and Nixon 2011; Yam et al. 2004) directly extract human body structure from the images with higher resolution images of a subject as well as higher computational cost. However, it is difficult to precisely estimate model parameters from low quality image sequences, which are captured from surveillance videos or crime scenes. Differently, the appearance-based methods (Han and Bhanu 2006; Iwama et al. 2012; Muramatsu et al. 2015; Sarkar et al. 2005; Sivapalan et al. 2013) mainly focus on extracting gait features from captured image sequences regardless of the underlying structure. Therefore, this kind of methods can perform recognition at lower resolutions, which makes them suitable for outdoor applications when the parameters of the body structure are difficult to be precisely estimated. Nonetheless, the human-crafted gait features in the existed methods can extremely hard to break through feature representation bottleneck when facing with the gait and appearance changes of a walking person with massive kinds of walking speed, viewpoint, clothing, and carrying.

Recently, because of the powerful feature learning abilities of deep neural networks, the data-driven deep learning methods achieve the state-of-the-art performance in plenty of fields (Gan et al. 2015; Hinton et al. 2012; Liu et al. 2015; Krizhevsky et al. 2012; Sutskever et al. 2014; Wang et al. 2015; Yan et al. 2017a, b). For example, the Convolution Neural Networks (CNN) can automatically learn commendable features from the given training images, which significantly improve the image classification accuracy. Therefore, some deep learning based gait recognition have been proposed to extract the robust gait feature (Castro et al. 2016; Feng et al. 2016; Wu et al. 2015; Zhang et al. 2016). CNN can automatically learn gait signatures from low-level motion features (i.e., optical flow components), which achieves the state-of-the-art performance in gait recognition. However, CNN has only learned the short-term motion features between adjacent frames rather than the periodic motion cue of gait sequences. Moreover, to learn sufficient features, the CNN requires a mass of training data for all the categories. Conversely, in the area of gait recognition, the number of subjects is very large (e.g., hundreds or thousands), with only a few examples per subject (Iwama et al. 2012; Ma et al. 2012; Sarkar et al. 2005; Yu et al. 2006). Besides, gait recognition for human identification is not a typical classification problem (Yuan et al. 2006; Zha et al. 2007). Therefore, we cannot directly use CNN on gait recognition as the huge domain gap between them.

Most of the widely-used gait recognition datasets provide gait energy images (GEI) (Han and Bhanu 2006), which are the average silhouettes along the temporal dimension.

Unlike other gait representations (Bobick and Johnson 2001) which consider gait as a sequence of templates (poses), GEI represents human motion sequence in a single image while preserving some temporal information. As an average template, GEI is not sensitive to incidental silhouette errors in individual frames. There are many alternatives for GEIs, e.g., chrono-gait images (Wang et al. 2010) and gait flow images (Lam et al. 2011). However, a recent empirical study by Iwama et al. (Iwama et al. 2012) shows that GEI, despite of its simplicity, is the most stable and effective kind of features for gait recognition on their proposed dataset with 4,007 subjects. Therefore, to solve the data limitation problem of deep neural network training in gait recognition, we use GEI instead of raw sequence of human gait in this paper as the input of the network. As removing most noisy information while keeping the major human shapes and body changes during walking, the GEI representation can help deep neural network quickly capture the discriminative biometrics information in human gait.

However, the GEI average one cycle of the human gait into one image, which undoubtedly loss most of the temporal information. This makes the GEI still cannot well handle the challenges of complexity surveillance environments, diverse subject-related factors, and the cross-view variance. As a motion pattern, the spatial and temporal dimensionality of human gait are both very important. The GEI models the spatial information in a gait cycle. To sufficiently model the temporal information, we can use the deep 3-dimensional convolutional networks (3D CNN) to learn the human gait Convolutional 3D (C3D) feature (Tran et al. 2015). The C3D can model appearance and motion information simultaneously and outperform the 2D CNN features on various video analysis tasks. The features extracted by these 3D CNN encapsulate information are generic, compact, simple and efficient. However, the original C3D features are trained by the video classification, which have huge domain gap with human gait recognition. In particular, aiming to further learn sufficient feature representations to tackle gait recognition for human identification, we exploit the 3D CNN based Siamese neural network (Bromley et al. 1993; Chopra et al. 2005), which can simultaneously minimize the distance between similar human gaits and maximize the distance between dissimilar pairs with a distance metric learning architecture.

Motivated by the above observations, and aiming to address the aforementioned gait recognition challenges, we proposed a Siamese neural network based gait recognition for human identification. First of all, to solve the data limitation problem, and holistically exploit the spatial-temporal information, we use GEI and C3D instead of raw sequence of human gait as gait feature. The GEI and C3D features are both extracted by a trained siamese neural network. With the well-learned gait spatial and temporal features, the Null Foley-Sammon Transform (NFST) is exploited to combine the two different features from spatial space and temporal space respectively. The NFST is one type of metric learning methods, whose goal is to learn a projection matrix W to map the different features into a latent metric null space. In the null space, the distances of features from the same object are much smaller than those of features from different objects. After that, the K-Nearest Neighbor (KNN) method (Muja and Lowe 2012; Liu et al. 2014) is exploited to identify the same person in surveillance environment with the combined features in null space. Finally, the evaluations on the world's largest and most comprehensive gait benchmark dataset demonstrate that the proposed method can impressively beyond state-of-the-art methods both in intra-view and inter-view gait-based human identification.

In contrast to our previous works [14], we propose a spatial-temporal Siamese Neural Network based gait recognition for human identification with GEI and C3D. The GEI chiefly represents the spatial information. By contrast, the C3D mainly represents the temporal information. To combine the spatial-temporal information together, we exploit the NFST to embed the GEI and C3D features into the null space to eliminate the domain gap. Compared with the previous system, we can achieve 17.28% Rank-1 accuracy improvement in the inter-view human identification. Moreover, we add the related work section to give more comprehensive literature survey and motivations from the perspective of a trade-off between current techniques. To evaluate the adaptation ability of the proposed gait recognition performance under different conditions, we conduct extensive experiments on two popular gait recognition datasets.

In summary, this paper makes the following contributions:

- In the end-to-end framework, we leverage the competitive GEI and C3D presentation as the input of network while holistically exploit the Siamese neural network to learn effective feature representations for human identification.
- We propose to exploit the NSFT to embed the GEI and C3D features into the null space to fuse the spatial-temporal information.
- The comprehensive evaluations show that we impressively outperform the state-of-the-art methods on the challenge gait benchmark dataset.

Related Work

As a burgeoning biometric identification technology, gait recognition attracts extensive attention due to its huge potential for automatic, remote accessed, robust, and

security human identification applications. A typical gait recognition framework contains five main components — 1) gait data acquisition; 2) gait data preprocessing; 3) gait cycle detection; 4) gait feature extraction; and 5) gait classification and matching. In this paper, we mainly focus on the last two steps. According to the feature presentation, we can classify the existing method into two categories: model based methods and appearance (i.e., mode-free) based methods. In this section, we review related research on gait feature extraction and gait matching, which are closely related to our work.

Gait Recognition

According to the feature presentation, we can classify the existing methods into two categories: model-based (Kusakunniran 2014; Lombardi et al. 2013; Nie et al. 2015; Sigal et al. 2012; Urtasun and Fua 2004; Wang et al. 2014; Wang et al. 2004; Yam et al. 2004) methods and appearance-based (i.e., mode-free) (Lam et al. 2011; Makihara et al. 2012, 2016; Muramatsu et al. 2015; Sarkar et al. 2005; Tao et al. 2007) methods.

Model-based approaches directly extract human body structure from gait sequences with higher resolution as well as higher computational cost. For example, Yam et al. designed two new models for modeling the leg motion of walking and running, which were used to extract gait feature (Yam et al. 2004). Urtasun et al. developed a tracking algorithm for identifying people and characterize their gait information (Urtasun and Fua 2004). Kusakunniran et al. proposed a new method to detect the Space-Time Interest Points (STIPs) where there are significant movements of human body along both spatial and temporal directions in local spatio-temporal volumes of a raw gait video (Kusakunniran 2014). Wang et al. proposed a visual recognition algorithm based upon fusion of static and dynamic body biometrics for gait recognition (Wang et al. 2004). In addition, precise estimating human pose is helpful for the feature extraction of gait images. Sigal et al. introduced a benchmark for measuring the performance of human motion estimation and designed a graphical model for representing the position and orientation of limbs (Sigal et al. 2012). Wang et al. proposed a method of estimating 3D human poses from a single image, which works in conjunction with an existing 2D pose detector (Wang et al. 2014). Nie et al. integrated action recognition and pose estimation into a framework, as the two tasks are closely related tasks for understanding human motion in videos (Nie et al. 2015). Zhang et al. presented a method to estimate a sequence of human poses in unconstrained videos, which could enforce the intra- and inter-frame body part constraints respectively without extra computational complexity (Zhang and Shah 2015). As model-based methods typically estimate the motion of the limbs, they are insensitive to viewpoint changes. However, it is extremely difficult to precisely calculate model parameters due to the low quality image sequences captured from surveillance videos in many situations.

On the other hand, appearance-based approaches extract gait features directly from captured image sequences. The major advantage is that they avoid the estimation of a latent model. We focus on appearance-based approaches in this paper. GEI (Han and Bhanu 2006) is an efficient spatio-temporal gait representation approach, which was proposed to characterize human walking properties for individual recognition by gait. Optical flow (Hu et al. 2013; Lam et al. 2011; Makihara et al. 2012) is widely adopted, which can provide a better avenue for representing gait body-shape invariant. For example, Makihara et al. constructed a smooth pseudo motion by using optical flows as silhouette motion features, then the smooth pseudo motion was utilized to match gaits from different people (Makihara et al. 2012). Lam et al. proposed a new gait representation, namely gait flow image (GFI), which further improved the accuracy of gait recognition (Lam et al. 2011). Hu et al. proposed a novel incremental framework based on optical flow, which could greatly improve the usability of gait traits in video surveillance applications (Hu et al. 2013). Other appearance-based methods are based on silhouette sequences (Makihara et al. 2006; Muramatsu et al. 2015; Sarkar et al. 2005; Tao et al. 2007). For example, Makihara et al. proposed a method of gait recognition from various view directions using frequency-domain features and a view transformation model (Makihara et al. 2006). Sarkar et al. introduced the HumanID gait challenge problem for providing a measure progress and characterizing the properties of gait recognition (Sarkar et al. 2005). Tao et al. focused on the representation and pre-processing of appearance-based models for human gait sequences and presented Gabor gait and tensor gait representations to further enhance their abilities for recognition tasks (Tao et al. 2007).

Recently, deep learning based gait recognition starts to replace traditional gait recognition. The Siamese convolutional neural network (SCNN) based gait recognition was proposed in Zhang et al. (2016) to automatically extract discriminative gait features for person re-identification. Feng et al. presented a novel feature learning method for gait recognition, which could extract body joint heatmap for each frame by exploiting pose estimation method and model the high level motion feature in the heatmap sequence (Feng et al. 2016). Shiraga et al. designed GEINet for gait recognition, which generated a set of similarities to the individual subjects (Shiraga et al. 2016). Wu et al. focused on capturing the co-occurrences and frequencies of features for better matching, and obtained excellent performances for

gait recognition (Wu et al. 2015). Castro et al. regarded person identification in video as gait recognition and learned high-level descriptors from low-level motion features (i.e., optical flow components) by using CNN (Castro et al. 2016). Wu et al. presented a CNN-based method via similarity learning by deep convolutional neural networks for gait recognition. It could automatically learn to recognize the most discriminative changes of gait features (Wu et al. 2017).

In summary, existing methods provide a avenue for gait recognition. However, they are extremely hard to capture the subtle periodic information of different gaits. In addition, recognition accuracy will be seriously affected when facing with the unaligned gait sequences in gait cycles.

3D Convolutional Neural Networks

3D CNN (Tran et al. 2015) allows the network to learn features from both the spatial and temporal dimensions, which can capture the motion information in video. 3D CNN have been successfully applied in action recognition (Cao et al. 2017; Ji et al. 2013; Varol et al. 2016; Tran et al. 2015; Yan et al. 2014), and action detection (Gao et al. 2017; Hou et al. 2017; Xu et al. 2017; Yan et al. 2014; Zolfaghari et al. 2017). For example, 3D CNN with small $3\times3\times3$ convolution kernels was employed for capturing the appearance and motion information encoded in video streams simultaneously (Gao et al. 2017; Ji et al. 2013; Tran et al. 2015). Cao et al. employed selective convolutional layer activations of 3D CNN to form a discriminative descriptor under the guidance of body joint positions, then the video feature was used for human action recognition (Cao et al. 2017). Xu et al. introduced a Region Convolutional 3D Network (R-C3D), which used a three-dimensional fully convolutional network to generate candidate temporal regions containing activities and classified selected regions into specific activities (Xu et al. 2017). Hou et al. proposed a Tube Convolutional Neural Network (T-CNN) which was an end-to-end deep network for action detection (Hou et al. 2017). It could recognize and localize action based on 3D convolution features in videos. Zolfaghari et al. designed a network architecture based on C3D network for action recognition that computed and integrated the most important visual cues : pose, motion, and the raw images (Zolfaghari et al. 2017).

In summary, compared to the above approaches, our proposed work aims to combine the spatial and temporal gait sequence information for gait recognition.

Deep Learning for Gait Recognition

We begin with going through the entire process of the gait feature learning framework. As illustrated in Fig. 2, the proposed approach for gait recognition consists following components. Firstly, we combine the raw sequence of surveillance images into GEIs, which are exploited as the input of the deep neural network. Instead of the conventional CNN, we use Siamese network, which can simultaneously minimize the distance between similar subjects and maximize the distance between dissimilar pairs, to learn sufficient spatial feature representations of gait for human identification. Secondly, the original walking sequences in one gait period are fed into the

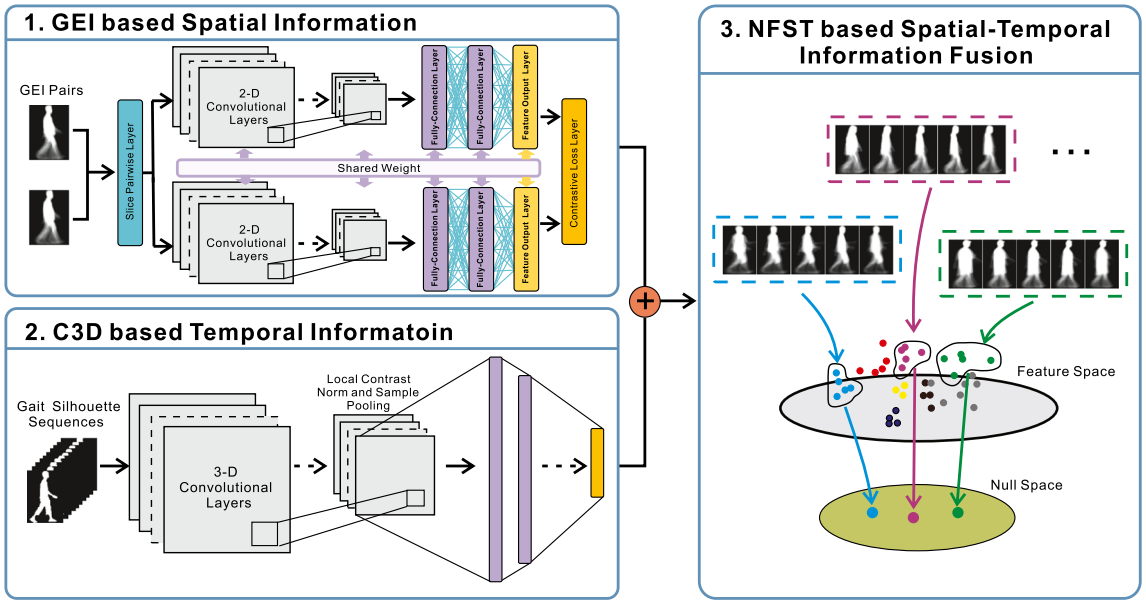

Fig. 2 The framework of the proposed spatial-temporal gait feature learning approach

fine-tuned C3D architecture to extract discriminative periodic temporal feature. Similarly, the Siamese neural network is utilized to learn sufficient temporal feature representations of human gait based on pairwise gait sequential frames. After that, the well-learned GEI-based spatial feature and C3D-based temporal feature are combined as the joint feature to represent human gait. Lastly, the concatenated spatiotemporal gait features are fused by NFST to be embedded into a discriminative latent null space. Finally, the final human recognition is made by comparing the Euclidean distances (i.e., matching scores) between the feature vectors of gallery and probe gait sequences in the null space. In the following section, we will describe each component of our method in detail.

GEI-based Spatial Information

For one sequence of human gait recorded by surveillance camera, we can extract the properly aligned silhouette of human (Boykov and Jolly 2001; Liu et al. 2013) and average them into the GEI representation. Here, we use GEI as it represents a human motion sequence in a single image while preserving part temporal information. In addition, this averaging operation cannot only well maintain the original information of the gait sequences, but also be robust to incidental silhouette errors in individual image. Therefore, GEI is popular applied in various gait analysis tasks (Hu et al. 2011; Iwama et al. 2012).

Some GEIs belonging to different subjects are shown in Fig. 3. When given a size-normalized and horizontal-aligned human walking binary silhouette sequence $I(x, y, t)$, the gray level GEI $G(x, y)$ is obtained by Eq. 1,

$$G(x, y) = \frac{1}{N} \sum_{t=1}^{N} I(x, y, t), \tag{1}$$

where N is the number of frames in one complete cycle of the sequence, t is the frame number of the sequence, x and y are values in the 2D image coordinate. After that, the computed GEIs are fed into our deep architecture to further learn the gait features.

Siamese Neural Network

The Siamese network was first introduced in Chopra et al. (2005) and Bromley et al. (1993) to be applied to face and signature verification tasks. The main idea of the network is to learn a function that maps input patterns into a latent space where similarity metric to be small for pairs of the same objects, and large for pairs from different ones. For the gait recognition task, we employ SCNN as the following reasons: 1) metric learning are capable of learning generic features useful for making predictions about unknown class distributions even when very few examples from these new distributions are available; and 2) Siamese network is best suited for verification scenarios where the number of classes is very large, and/or examples of all the classes are not available at the time of training. Definitely, gait recognition is one of such verification scenarios.

Network Architecture for Pairwise GEI Training As shown in Fig. 4, the Siamese neural network designed for GEI-based spatial feature training contains two parallel CNN architectures, which of them consist of two parts: 1) two convolution layers and max-pooling layers, and 2) three fully connection layers. This network accepts inputs of size 128×88×3 pixels. Using shorthand notation, the full architecture of each branch is $C(20, 5, 1)$-N-P-$C(50, 5, 1)$-N-P-$FC(500)$-$FC(10)$-$FC(2)$, where $C(d, f, s)$ indicates a convolutional layer with d filters of spatial size $f \times f$, applied to the input with stride s. $FC(n)$ is a fully-connected layer with n nodes. All max-pooling layers P pool spatially in non-overlapping 2×2 regions and all normalization layers N use the same parameters: $n = 5$, $alpha = 10^{-4}$, and $beta = 0.75$. The final feature output layers are connected to a contrastive loss layer.

Contrastive Loss For gait recognition, we want to learn a nonlinearly function which maps gait sequences to points in

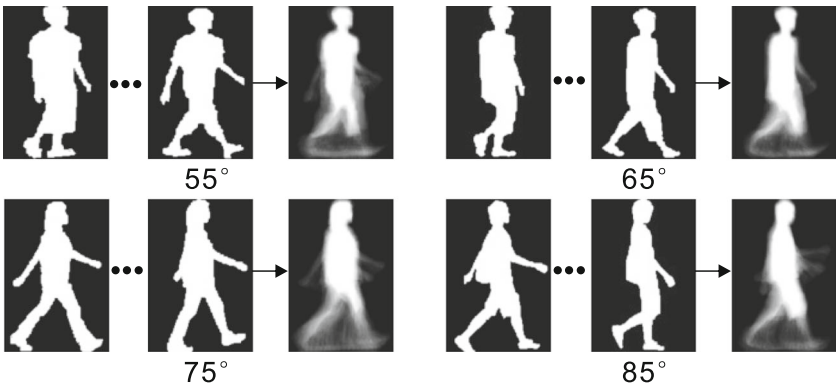

Fig. 3 Some GEIs of different subjects in the OU-ISIR LP dataset (Iwama et al. 2012) in terms of view point 55°, 65°, 75°, and 85°

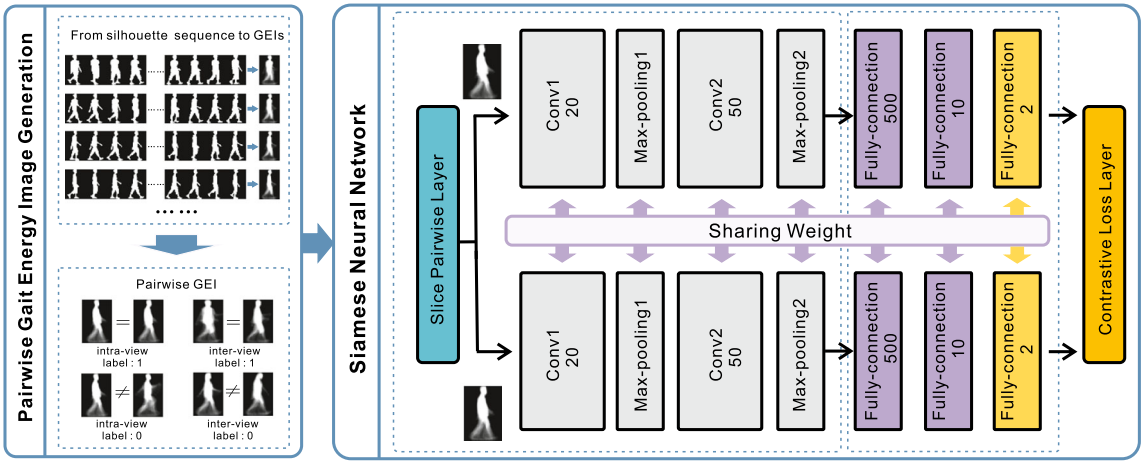

Fig. 4 The architecture of GEI-based spatial feature learning framework with Siamese neural network

a low dimensional space. Moreover, it makes positive pairs close enough, whereas negative pairs are far away at least by a margin. Therefore, we apply the contrastive loss layer to connect the two branches in the network.

Mathematically, considering a pair of GEIs x_1 and x_2, let y be a binary label of the pair, $y = 1$ if the images x_1 and x_2 belong to the same subject (i.e., "genuine pair") and $y = 0$ otherwise (i.e., "impostor pair"). W is the shared parameter matrix throughout the Siamese architecture which needs to be learned. We can use W to map x_1 and x_2 into $S_W(x_1)$ and $S_W(x_2)$, which are the two points in the latent low-dimensional space. Then the distance $E_W(x_1, x_2)$ between x_1 and x_2 can be measured by:

$$E_W(x_1, x_2) = ||S_W(x_1) - S_W(x_2)||_2^2. \tag{2}$$

We can define the contrastive loss function as follows:

$$\mathcal{L}(W) = \sum_{i=1}^{P} L(W, (y, x_1, x_2)^i), \tag{3}$$

$$\begin{aligned} L(W, (y, x_1, x_2)^i) = {} & (1 - y) \cdot max(m - E_W(x_1, x_2)^i, 0) \\ & + y \cdot E_W(x_1, x_2)^i, \end{aligned} \tag{4}$$

where $(y, x_1, x_2)^i$ is the i-th pair, which is composed of a pair of GEIs with corresponding label y, P is the number of the training pairs. The positive number m can be interpreted as margin. Then the total loss is the sum of all gait sequence pair losses.

Network Training We consider the gait recognition as binary classification. Training data includes gait sequence pairs and label. In the implementation, the training set is selected from OULP-C1V1-A-Gallery (Iwama et al. 2012) dataset, with 20,000 similar GEI pairs and randomly selected 20,000 dissimilar pairs. In the training stage, the two branches of the network will be optimized simultaneously with the weight sharing mechanism. Pairwise images with similar or dissimilar labels separately entrance the two CNNs. Then the output of the CNNs are combined by the contrastive layers to compute the contrastive loss. After that, the back-propagating with contrastive loss is used to fine-tune the model.

Our training algorithm adopts the mini-batch stochastic gradient descent for optimizing the objective function. The training data is divided into mini-batches. Training errors are calculated upon each mini-batch in the contrastive loss layer and backward propagated to the lower layers, and network weights are updated simultaneously. We use the "step" learning rate policy. We initialize the learning rate to 10^{-4} and gamma to 10^{-1}. Momentum and weight decay are set to 0.9 and 0.0005 respectively. The only tuned parameters are the hidden vector size, learning rate and margin for the contrastive loss function (see Eq. 4). As positive pairs and negative pairs have different data distribution, they can bring about data imbalance and overfitting. To avoid these problems, we randomly dropout 50% neurons of the last fully-connected layer in the training process. With more rounds over the training data, the model is trained until it converges.

Spatial Feature Extraction In the feature extraction stage, we discard the Siamese architecture and contrastive loss layer. We send the query GEI into one of the CNNs, then compute the feed-forward network based on the matrix multiplication for one time to extract features. The whole scheme will be very efficient. The spatial information of all gallery and probe gait sequences are extracted and mapped using the above feature extractor. In the following experiments, we denote the spatial feature representation as "GEI.SCNN".

C3D-based Temporal Information

In this subsection, we elaborate steps for learning and extracting C3D based temporal gait features. As shown in Fig. 5, we firstly try to extract motion features from the input gait video by fine-tuning the 3D CNN on the pre-trained C3D model. Next, we design a Siamese neural network specifically for the above C3D feature for automatically learning the periodic features of gait sequences. Finally, with the well-trained SCNN, we can extract the final C3D-based temporal information for better human identification.

Fine-Tuning Classifier for Capturing Sequential Information 3D CNN (Ji et al. 2013; Tran et al. 2015) is well-suited for spatiotemporal gait feature learning because 1) in 3D CNN, convolution and pooling operations are performed spatio-temporally. Conversely, in 2D CNN they are done only in spatial space, which lacks ability to capture temporal information from continued frames; and 2) the periodic motion is the most significant difference between the gait recognition and the traditional action recognition task while the procedure of GEI's computation is often accompanied by the loss of temporal and periodic information in gait sequences. Hence, only 3D CNN preserves the temporal information of the input signals resulting in an output volume. Therefore, we adopt the C3D architecture as described in Tran et al. (2015) to model the temporal motion and extract more discriminative temporal features for better gait recognition.

To capture the motion information of gait sequences, we select the C3D-based ResNet-18 (He et al. 2016), which did exceptionally well in video classification, and fine-tune from their learned representations on the currently largest video classification benchmark: Sports-1M dataset (Karpathy et al. 2014). We only change the 487 label output in Sport-1M categories to the number of subjects in the gait dataset. Specifically, inspired by Tran et al. (2015), we split every gait video into 16-frame clips with an overlap of 8 frames. Clips are resized to have a frame size of $128\times171\times3$ pixels. In the fine-tuning stage, we randomly crop input clips into $128\times171\times16$ crops for spatial and temporal filtering. We also horizontally flip them with 50% probability. Fine-tuning is done by SGD with mini-batch size of 50 examples. Initial learning rate is 0.001, and is divided by 2 every 150K iterations. The optimization is stopped at 1.9M iterations. After that, we extract C3D features: pooling5 for each clip. The features for one subject's gait sequence are computed by averaging the clip features separately, followed by an L_2 normalization. The fine-tuned model can efficiently capture the discriminative features of human gait frames, which are robust to viewpoint changes. The features are used as the input of the SCNN for further training. Consequently, we will process each gait sequence with the fine-tuned ResNet-18 model to obtain gait sequence feature vectors.

C3D Feature Learning based on SCNN After obtaining C3D features for each gait sequence, the target for gait recognition is to retrieve a set of matching gallery gait sequences features for a given query gait sequence feature. In particular, in the field of gait recognition, the number of categories can be very large, with only a few examples per category. Therefore, we design a Siamese neural network for learning subtle C3D motion features. Using shorthand notation, the full architecture of each parallel network is $FC(512)$-$FC(256)$-$FC(128)$-$FC(64)$-$dropout(0.7)$-$Output(2)$, where $FC(n)$ are fully-connected layers with n nodes and $Output(2)$ are the feature output layers which are connected to a contrastive loss layer.

Our proposed SCNN for C3D feature learning takes pairs of gait sequences feature vectors as inputs, and maps each feature vector into a latent feature space, which are then compared using Euclidean distance. In the implementation, we randomly select two sequences belonging to the same identity as positive pairs, whereas we randomly select two

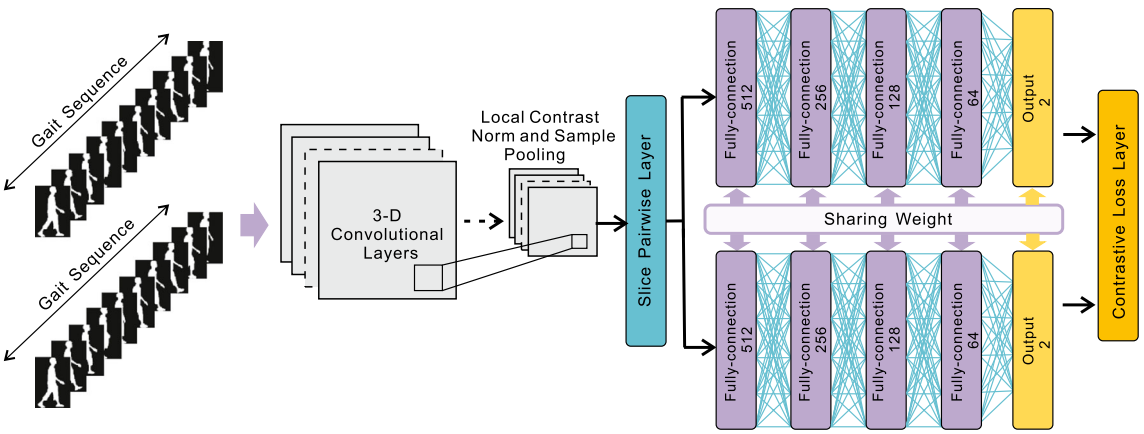

Fig. 5 The architecture of C3D-based temporal feature learning framework with Siamese neural network

sequences belonging to the different identities as negative pairs. Specifically, given a gait sequence pair (s_1, s_2), let $y \in [0, 1]$ be a binary label of the pair. We denote $y = 1$ for indicating that s_1 and s_2 belong to the same identity and $y = 0$ for indicating that they belong to the different identities. Following the "Siamese Neural Network", we use all the positive pairs and randomly select the same number of negative pairs as training data for training our network.

Temporal Feature Extraction In the feature extraction stage, given a probe/gallery gait sequence, we firstly extract the motion feature vector from the fine-tuned C3D model. Then, we feed the feature vector into one of the network from SCNN, and compute the feed-forward network for one time to convert the original C3D feature into a new latent space. In the following experiments, we evaluate both feature representations from $FC(512)$ and $FC(256)$, which give better performance comparing with the activation from other fully-connected layers. We denote the temporal feature representation as "C3D.SCNN" in our experiments.

NFST-based Spatial-Temporal Information Fusion

A simple scheme to fuse the GEI-based spatial feature and the C3D-based temporal feature is directly concatenating as an unified feature vector. However, it cannot effectively integrate the complementary spatial and temporal information and is extremely hard to obtain the view-invariant description which occurs frequently in the real world human identification scenario. The NFST was first introduced to address the small sample size problem in face recognition (Guo et al. 2006). Zhang et al. (2016) proposed a Kernalized NFST for person re-identification by mapping the multiple features into a discriminative null space and significantly outperforms the state-of-the-art methods.

In this paper, we propose a Null space based Spatial-Temporal gait feature (NuSTG) to extract effective and robust representation for gait recognition. The proposed NFST fusion method can take advantage of the two kinds of gait features, which respectively capture different aspects of information from gait sequences, i.e., spatial and temporal features. Firstly, these two types of features, F_s and F_t, are extracted from the all probe and gallery gait videos and concatenated to obtain the original spatiotemporal feature as $F = [F_s, F_t]$. Formally, the training features F are kernelized by kernal function $\Psi(\cdot)$ to obtain $\Psi(F)$. Then, the projection matrix W of discriminative null space is learned by NFST on $\Psi(F)$ as in Zhang et al. (2016). We aim to learn the optimal projection matrix W so that each column, denoted as w, is an optimal discriminant direction that maximizes the Fisher discriminant criterion:

$$J(\mathrm{w}) = \frac{\mathrm{w}^\top S_b \mathrm{w}}{\mathrm{w}^\top S_w \mathrm{w}}, \tag{5}$$

where w denotes each column in W, S_b is the inter-class scatter matrix and S_w is the intra-class scatter matrix. It meets the following two constrains, i.e., zero intra-class scatter and positive inter-class scatter:

$$\mathrm{w}^\top S_w \mathrm{w} = 0, \tag{6}$$

$$\mathrm{w}^\top S_b \mathrm{w} > 0. \tag{7}$$

As shown in Fig. 2, with W, we can learn a discriminative null space where the training data of the same identities are collapsed into one single point, resulting in C points in the space, where C is the number of categories. The original features F can be mapped into a latent metric space in which the distances of features from the same object are much smaller than those of features from different objects.

Experiments

Dataset and Experimental Setting

In our experiments, we evaluate the proposed gait recognition framework on the OULP-C1V1-A dataset from the OU-ISIR LP gait benchmark (Iwama et al. 2012). Recently, there are several large-scale datasets (Sarkar et al. 2005; Yu et al. 2006) have been built for gait-based human identification or attribute classification. Compared with above datasets, OU-ISIR LP dataset has the following properties which make it a valuable and challenging dataset for gait-based human identification. First of all, OU-ISIR LP contains largest number of subjects with a wide age and an almost balanced gender distribution in the world. Moreover, the dataset records two sequences for each subject: probe (i.e., query) sequence and gallery (i.e., source) sequence. Such standard directory structure of gallery and probe offers test bed for fairly algorithm comparison. The sequences are further constituted by silhouette images, which are normalized into 128×88 pixels. Some examples of the dataset can be found in Fig. 3. In addition, each sequence is divided into 4 slices based on the observation angles (55°, 65°, 75°, and 85°), which makes the dataset can be used for both inter-view and intra-view human identification task.

In this paper, we firstly compare different gait-based methods on human identification within the same observation angle (i.e., intra-view human identification). Then we evaluate state-of-the-art algorithms on inter-view human identification task, which is more challenging. The criteria for evaluating the methods is the rank-1 and rank-5 identification rates, which denote the percentages of correct subjects out of all the subjects appearing within the first and fifth ranks, respectively.

Evaluation on Intra-View Human Identification

To begin with, we evaluate nine gait recognition approaches for intra-view identification task. The details of approaches are described as follows.

1. **GEI Template**. The conventional GEI based template matching strategy (Iwama et al. 2012).
2. **FDF Template**. FDF is generated by applying a Discrete Fourier Transform of the temporal axis to the silhouette images in one gait cycle (Iwama et al. 2012).
3. **HWLD**. It is the histogram of weighted local directions extracted from the GEIs (Sivapalan et al. 2013). Note, the authors of Sivapalan et al. (2013) only tested their HWLD method in near-profile view set (A-85).
4. **GEI.fine-tuned**. It is conventional 2D CNN-based feature learning method fine-tuned from the pre-trained AlexNet network (Krizhevsky et al. 2012). As in Zhang et al. (2016), GEI.fine-tuned took the activations from the first fully connected layers as feature representations to identify the same person.
5. **GEI.SCNN**. As the previous work introduced in Zhang et al. (2016), GEI.SCNN method adopted Siamese convolutional neural network as feature extractor for GEIs. The output activation of the first fully-connected layer (500-D) is extracted as the GEI-based spatial feature of input gait sequence.
6. **C3D.fine-tuned**. To evaluate original temporal features extracted by 3D CNN, we directly use the ResNet-18 (He et al. 2016) network to fine-tune the pre-trained C3D model on the Sport-1M dataset (Karpathy et al. 2014). As discussed in "C3D-based Temporal Information" , the fine-tuned C3D model is employed as a feature extractor for high-level attributes. At last, we obtain a 512-D feature from the pooling5 layer of the neural network to represent the temporal feature of gait sequence.
7. **C3D.SCNN**. After obtaining C3D.fine-tuned feature vector, we attempt to train a Siamese neural network based on C3D feature. Different from GEI-based SCNN, C3D-based SCNN only consists two branches of MLP, but without convolutional layers. We discussed details of the network architecture in "C3D-based Temporal Information". During testing, the feature is extracted from the first fully-connected layer as a 512-D vector.
8. **STG (GEI.SCNN+C3D.SCNN)**. We directly concatenate the GEI.SCNN based spatial feature (500-D) and C3D.SCNN temporal feature (512-D) to obtain the Spatial-Temporal Gait (STG) feature vector (1012-D) for gait-based human identification task.
9. **NuSTG (GEI.SCNN+C3D.SCNN)**. The proposed method for training Spatial-Temporal Gait features which concatenates the GEI.SCNN feature and C3D.SCNN feature within the Null space by NFST. We call this method NuSTG.

Table 1 shows the comparison results of our approach and state-of-the-art gait recognition methods. As we can see, the deep learning based method is better than hand-crafted feature based techniques in all test cases and achieves the state-of-the-art performance. Compared to GEI, FDF and HWLD, spatial and temporal information are able to automatically learn commendable features from the given GEIs and gait sequences, which obviously improve the identification rate. Furthermore, compared with traditional CNN-based method, our Siamese network based method also achieves much better accuracy because its distance metric learning architecture can well solve the verification scenario of gait recognition based human identification. Some visual example results of the proposed framework on the OULP-C1V1-A dataset are shown in Fig. 6. From these examples, cascade convolutional architecture of Siamese network has ability to capture the massive complexity structure around several ambiguous regions such as human neck, hip and shoulder. Compared with spatial and temporal feature, we can see that on rank-1 results, the spatial feature is better. Conversely, the temporal feature is better on rank-5 results. Compared with two different fusion methods, the STG is similar with spatial features, which cannot well combine the spatial-temporal information. Finally, the null space based fusion method NuSTG achieves comparable results on all the degrees, which demonstrates that this fusion strategy can holistically exploit the complementary nature of spatial-temporal information.

Nonetheless, compared with inter-view human identification, the intra-view cases are easier. All the proposed methods can well solve this problem. Next, we will compare all methods on the inter-view human identification, which can observably demonstrate the power of the proposed NFST based fusion method.

Evaluation on Inter-View Human Identification

For gait-based human identification, the changes in view between query gait sequences and source samples occurs frequently in the real-world human identification scenario. Therefore, we further verify the robustness of our proposed methods in inter-view gait recognition task. Here is the details of different state-of-the-art inter-view gait recognition methods.

1. **AVTM**. The Arbitrary View Transformation Model (AVTM) was particularly designed for gait-based human identification with inter-view matching by applying an extra 3D gait volume for training view transformation model. The AVTM method had several

Table 1 Performance comparison of different methods in term of the rank-1 and rank-5 identification rates on intra-view gait recognition. The bold values indicate the best accuracy, and the hyphens (--) denote out of scope

Methods	Rank-1 Recognition Rate (%)				Rank-5 Recognition Rate (%)			
	A-55	A-65	A-75	A-85	A-55	A-65	A-75	A-85
Baseline:								
GEI template (Iwama et al. 2012)	84.70	86.63	86.91	85.72	92.39	92.84	92.78	93.01
FDF template (Iwama et al. 2012)	83.89	85.49	86.59	85.90	91.53	92.81	92.88	92.83
HWLD (Sivapalan et al. 2013)	--	--	--	87.70	--	--	--	94.70
Spatial feature:								
GEI.fine-tuned (Zhang et al. 2016)	73.96	76.71	77.87	78.82	86.64	88.67	89.39	90.09
GEI.SCNN (Zhang et al. 2016)	**90.12**	**91.14**	**91.18**	**90.43**	**94.98**	95.90	95.92	95.97
Temporal feature:								
C3D.fine-tuned	79.95	84.67	85.95	89.07	93.74	95.25	96.11	96.98
C3D.SCNN	82.65	86.63	87.34	90.37	94.63	**95.92**	**96.67**	**96.98**
Fusion:								
STG	82.68	86.63	87.36	90.37	94.63	**95.92**	**96.67**	**96.98**
NuSTG	85.21	89.36	89.66	90.03	92.66	95.15	95.25	95.60

extended versions including AVTM_PdVS (Muramatsu et al. 2015), AVTM (Muramatsu et al. 2015) and woVTM (Muramatsu et al. 2015).

2. **RankSVM**. RankSVM (Martín-Félez and Xiang 2012) was a typical representative metric learning based approach aiming at gait recognition task with covariate variations especially for inter-view matching.

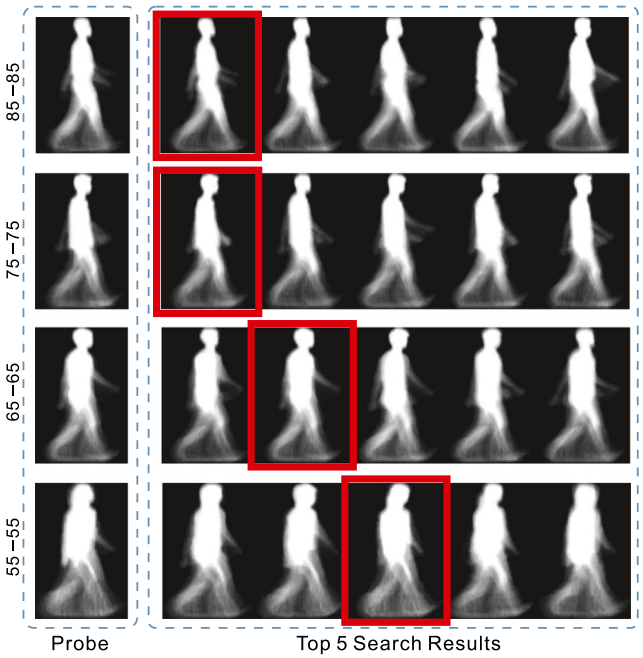

Fig. 6 Visual searching results on intra-view human identification task. Red solid bounding box indicates true match

3. **GEI.SCNN**. As introduced before, GEI.SCNN (Zhang et al. 2016) method adopted Siamese convolutional neural network as feature extractor for GEIs which also can well handle the inter-view human identification problem.

As shown in Table 2, the performance improvement of our framework is consistent and stable, i.e., all the inter-view test cases are improved compared to other methods. Note that, the AVTM and RankSVM methods only selected 1,912 subjects in OU-ISIR LP dataset, whose data were captured for evaluation by calibrated cameras for testing. Differently, we evaluate our methods on the whole 3,835 persons set, which is more difficult. The results demonstrate that the proposed method is quite robust to view-change variations. The reason is that in the training stage, we send both intra- and inter-view pairwise gait samples into the Siamese neural network. In this way, we can train the similarity metric to be small for pairs from same subjects, and large for pairs from different subjects, which enhance the robustness of the gait-based human identification method under the view-change condition. Compared with spatial features, the temporal features achieve much better performance on the inter-view human identification. The results demonstrate that compared with temporal information, spatial information are more sensitive to the view point changes. Differently, as the temporal features mainly take advantage of motion information, it is more robust. Moreover, the proposed NFST based spatial-temporal fusion method significantly outperforms all the other methods, especially better than the early fusion strategy. The results show that the metric

Table 2 Performance comparison of different methods in term of the rank-1 identification rates on cross-view gait recognition. The bold values indicate the best accuracy, and the hyphens (--) denote out of scope

Probe	Gallery	A-55	A-65	A-75	A-85
A-55	AVTM_PdVS (Muramatsu et al. 2015)	--	76.20	61.45	45.50
	AVTM (Muramatsu et al. 2015)	--	77.72	64.54	42.69
	woVTM (Muramatsu et al. 2015)	--	53.61	14.28	8.94
	RankSVM (Martín-Félez and Xiang 2012)	--	--	19.98	12.60
	GEI.SCNN (Zhang et al. 2016)	--	65.76	32.92	19.48
	C3D.fine-tuned	--	78.25	71.50	64.02
	C3D.SCNN	--	81.22	75.10	67.41
	STG	--	81.22	75.10	67.44
	NuSTG	--	**87.80**	**87.39**	**87.20**
A-65	AVTM_PdVS (Muramatsu et al. 2015)	75.99	--	77.09	65.48
	AVTM (Muramatsu et al. 2015)	75.63	--	76.36	62.76
	woVTM (Muramatsu et al. 2015)	49.84	--	68.62	38.18
	RankSVM (Martín-Félez and Xiang 2012)	--	--	62.71	36.66
	GEI.SCNN (Zhang et al. 2016)	72.58	--	78.54	51.83
	C3D.fine-tuned	77.44	--	81.31	77.01
	C3D.SCNN	79.68	--	84.24	79.84
	STG	79.68	--	84.24	79.87
	NuSTG	**84.84**	--	**88.99**	**88.83**
A-75	AVTM_PdVS (Muramatsu et al. 2015)	60.25	76.20	--	76.52
	AVTM (Muramatsu et al. 2015)	59.88	74.90	--	76.31
	woVTM (Muramatsu et al. 2015)	13.54	67.63	--	77.72
	RankSVM (Martín-Félez and Xiang 2012)	15.49	58.47	--	60.83
	GEI.SCNN (Zhang et al. 2016)	39.13	78.30	--	81.22
	C3D.fine-tuned	70.53	81.19	--	85.29
	C3D.SCNN	72.37	83.61	--	85.84
	STG	72.37	83.63	--	85.84
	NuSTG	**83.89**	**88.30**	--	**89.54**
A-85	AVTM_PdVS (Iwama et al. 2012)	40.48	60.62	73.12	--
	AVTM (Iwama et al. 2012)	40.17	61.87	74.32	--
	woVTM (Iwama et al. 2012)	7.16	33.42	77.14	--
	RankSVM (Martín-Félez and Xiang 2012)	10.41	34.41	57.79	--
	GEI.SCNN (Zhang et al. 2016)	19.45	44.93	71.39	--
	C3D.fine-tuned	59.42	68.86	74.09	--
	C3D.SCNN	60.71	70.08	75.82	--
	STG	60.71	70.11	75.82	--
	NuSTG	**74.04**	**77.75**	**79.02**	--

learning can markedly make the distances of features from the same object become much smaller than those of features from different objects in the new null space.

Conclusions

In this paper, we have investigated on leveraging deep learning method to extract robust and discriminative spatial-temporal gait feature for human identification. In the end-to-end framework, we utilize the competitive GEI and C3D presentation as the input of network to capture the major human shapes and body changes during walking. Furthermore, we holistically exploit the Siamese neural network to learn effective gait features which directly computs the similarity between two human gaits with parallel neural network architecture. More important, we propose to use the NFST to combine the GEI and C3D features into the null space to learn more comprehensive spatial-temporal gait information. The experimental results on the benchmark dataset demonstrate the effectiveness and efficient of our proposed method. In the future, we will try

to combine the gait data acquisition, preprocessing, cycle detection, feature extraction, classification and matching into one end-to-end network.

Information Sharing Statement

In the article, the deep learning networks are implemented through Caffe (http://caffe.berkeleyvision.org/). We worked with data from the following data sources: OULP-C1V1-A dataset (http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitLP.html).

Acknowledgements This work is partially supported by the Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (No. 61720106007), the NSFC-Guangdong Joint Fund (No. U1501254), the National Natural Science Foundation of China (No. 61602049), and the Cosponsored Project of Beijing Committee of Education.

References

Ariyanto, G., & Nixon, M.S. (2011). Model based 3d gait biometrics. In *Proceedings of international joint conference on biometrics, pp. 1–7. IEEE*.

Bobick, A.F., & Johnson, A.Y. (2001). Gait recognition using static, activity-specific parameters. In *Proceedings of IEEE conference on computer vision and pattern recognition, vol. 1, pp. I–I. IEEE*.

Boykov, Y., & Jolly, M. (2001). Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings of international conference on computer vision, vol. 1, pp. 105–112. IEEE*.

Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R. (1993). Signature verification using a siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*(04), 669–688.

Cao, C., Zhang, Y., Zhang, C., Lu, H. (2017). Body joint guided 3d deep convolutional descriptors for action recognition. CoRR arXiv:1704.07160.

Castro, F.M., Marín-Jiménez, M.J., Guil, N., de la Blanca, N.P. (2016). Automatic learning of gait signatures for people identification. CoRR arXiv:1603.01006.

Chen, Z., Ngo, C., Zhang, W., Cao, J., Jiang, Y. (2014). Name-face association in web videos: A large-scale dataset, baselines, and open issues. *J. Comput. Sci. Technol*, *29*(5), 785–798.

Chen, Z., Zhang, W., Deng, B., Xie, H., Gu, X. (2017). Name-face association with web facial image supervision. Multimedia Systems (4), 1–20.

Chopra, S., Hadsell, R., LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of IEEE conference on computer vision and pattern recognition, vol. 1, pp. 539–546. IEEE*.

Feng, Y., Li, Y., Luo, J. (2016). Learning effective gait features using lstm. In *23rd international conference on pattern recognition, pp. 325–330. IEEE*.

Gan, C., Wang, N., Yang, Y., Yeung, D., Hauptmann, A.G. (2015). Devnet: A deep event network for multimedia event detection and evidence recounting. In *IEEE conference on computer vision and pattern recognition, pp. 2568–2577*.

Gao, J., Yang, Z., Sun, C., Chen, K., Nevatia, R. (2017). TURN TAP: temporal unit regression network for temporal action proposals. CoRR arXiv:1703.06189.

Guo, Y.F., Wu, L., Lu, H., Feng, Z., Xue, X. (2006). Null foley–sammon transform. *Pattern recognition*, *39*(11), 2248–2251.

Han, J., & Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(2), 316–322.

He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778*.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Hou, R., Chen, C., Shah, M. (2017). Tube convolutional neural network (T-CNN) for action detection in videos. CoRR arXiv:1703.10664.

Hu, M., Wang, Y., Zhang, Z., Zhang, D. (2011). Gait-based gender classification using mixed conditional random field. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, *41*(5), 1429–1439.

Hu, M., Wang, Y., Zhang, Z., Zhang, D., Little, J.J. (2013). Incremental learning for video-based gait recognition with LBP flow. *IEEE Transactions Cybernetics*, *43*(1), 77–89.

Iwama, H., Okumura, M., Makihara, Y., Yagi, Y. (2012). The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. *IEEE Transactions on Information Forensics and Security*, *7*(5), 1511–1521.

Ji, S., Xu, W., Yang, M. (2013). Yu, K.: 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 221–231.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.F. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of IEEE conference on computer vision and pattern recognition, pp. 1725–1732. IEEE*.

Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems, pp. 1097–1105*.

Kusakunniran, W. (2014). Attribute-based learning for gait recognition using spatio-temporal interest points. *Image Vision Comput*, *32*(12), 1117–1126.

Lam, T.H.W., Cheung, K.H., Liu, J.N.K. (2011). Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognition*, *44*(4), 973–987.

Liu, W., Mei, T., Zhang, Y. (2014). Instant mobile video search with layered audio-video indexing and progressive transmission. *IEEE Transactions on Multimedia*, *16*(8), 2242–2255.

Liu, W., Mei, T., Zhang, Y., Che, C., Luo, J. (2015). Multi-task deep visual-semantic embedding for video thumbnail selection. In *IEEE conference on computer vision and pattern recognition, pp. 3707–3715*.

Liu, W., Zhang, Y., Tang, S., Tang, J., Hong, R., Li, J. (2013). Accurate estimation of human body orientation from RGB-D sensors. *IEEE Transactions on Cybernetics*, *43*(5), 1442–1452.

Lombardi, S., Nishino, K., Makihara, Y., Yagi, Y. (2013). Two-point gait: decoupling gait from body shape. In *IEEE international conference on computer vision, pp. 1041–1048*.

Ma, H., & Liu, W. (2017). Progressive search paradigm for internet of things. IEEE Multimedia. https://doi.org/10.1109/MMUL.2017.265091429.

Ma, H., Zeng, C., Ling, C.X. (2012). A reliable people counting system via multiple cameras. *ACM Transaction on Intelligent Systems and Technology*, *3*(2), 31.

Makihara, Y., Rossa, B.S., Yagi, Y. (2012). Gait recognition using images of oriented smooth pseudo motion. In *Proceedings of the IEEE international conference on systems, Man, and Cybernetics, SMC 2012, Seoul, Korea (South), October 14-17, 2012, pp. 1309–1314*.

Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., Yagi, Y. (2006). Gait recognition using a view transformation model in the frequency domain. In *Proceedings of european conference on computer vision, pp. 151–163*.

Mannini, A., Trojaniello, D., Cereatti, A., Sabatini, A.M. (2016). A machine learning framework for gait classification using inertial sensors: Application to elderly, post-stroke and huntington's disease patients. *Sensors*, *16*(1), 134.

Martín-Félez, R., & Xiang, T. (2012). Gait recognition by ranking. In *Proceedings of european conference on computer vision, pp. 328–341. Springer*.

Muja, M., & Lowe, D.G. (2012). Fast matching of binary features. In *Proceedings of computer and robot vision, pp. 404–410*.

Muramatsu, D., Shiraishi, A., Makihara, Y., Uddin, M., Yagi, Y. (2015). Gait-based person recognition using arbitrary view transformation model. *IEEE Transactions on Image Processing*, *24*(1), 140–154.

Nie, B.X., Xiong, C., Zhu, S. (2015). Joint action recognition and pose estimation from video. In *Proceedings of IEEE conference on computer vision and pattern recognition, pp. 1293–1301*.

Ren, P., Tang, S., Fang, F., Luo, L., Xu, L., Bringas-Vega, M.L., Yao, D., Kendrick, K.M., Valdes-Sosa, P.A. (2017). Gait rhythm fluctuation analysis for neurodegenerative diseases by empirical mode decomposition. *IEEE Transactions Biomed. Engineering*, *64*(1), 52–60.

Samà, A., Pèrez-Lòpez, C., Martín, D.R., Català, A., Aròstegui, J.M., Cabestany, J., de Mingo, E., Rodríguez-Molinero, A. (2017). Estimating bradykinesia severity in parkinson's disease by analysing gait through a waist-worn sensor. *Comp. in Bio. and Med.*, *84*, 114–123.

Sarkar, S., Phillips, P.J., Liu, Z., Vega, I.R., Grother, P., Bowyer, K.W. (2005). The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions Pattern Anal. Mach. Intell*, *27*(2), 162–177.

Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y. (2016). Geinet: View-invariant gait recognition using a convolutional neural network. In *Proceedings of international conference on biometrics, pp. 1–8*.

Sigal, L., Isard, M., Haussecker, H.W., Black, M.J. (2012). Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision*, *98*(1), 15–48.

Sivapalan, S., Chen, D., Denman, S., Sridharan, S., Fookes, C. (2013). Histogram of weighted local directions for gait recognition. In *Proceedings of computer vision and pattern recognition workshop, pp. 125–130. IEEE*.

Sutskever, I., Vinyals, O., Le, Q.V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems, pp. 3104–3112*.

Tao, D., Li, X., Wu, X., Maybank, S.J. (2007). General tensor discriminant analysis and gabor features for gait recognition. *IEEE Transactions Pattern Anal. Mach. Intell*, *29*(10), 1700–1715.

Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of international conference on computer vision, pp. 4489–4497*.

Urtasun, R., & Fua, P. (2004). 3d tracking for gait characterization and recognition. In *Proceedings of 6th IEEE international conference on automatic face and gesture recognition, pp. 17–22*.

Varol, G., Laptev, I., Schmid, C. (2016). Long-term temporal convolutions for action recognition. CoRR arXiv:1604.04494.

Wang, B., Tang, S., Zhao, R., Liu, W., Cen, Y. (2015). Pedestrian detection based on region proposal fusion. In *Proceedings of international workshop on multimedia signal processing, pp. 1–6. IEEE*.

Wang, C., Wang, Y., Lin, Z., Yuille, A.L., Gao, W. (2014). Robust estimation of 3d human poses from a single image. In *Proceedings of IEEE conference on computer vision and pattern recognition, pp. 2369–2376*.

Wang, C., Zhang, J., Pu, J., Yuan, X., Wang, L. (2010). Chrono-gait image: A novel temporal template for gait recognition. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I, pp. 257–270*.

Wang, L., Ning, H., Tan, T., Hu, W. (2004). Fusion of static and dynamic body biometrics for gait recognition. *IEEE Transactions Circuits Syst. Video Techn*, *14*(2), 149–158.

Wang, L., Tan, T., Ning, H., Hu, W. (2003). Silhouette analysis-based gait recognition for human identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(12), 1505–1518.

Wu, Z., Huang, Y., Wang, L. (2015). Learning representative deep features for image set analysis. *IEEE Transactions Multimedia*, *17*(11), 1960–1968.

Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T. (2017). A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(2), 209–226.

Xia, Y., Gao, Q., Ye, Q. (2015). Classification of gait rhythm signals between patients with neuro-degenerative diseases and normal subjects: Experiments with statistical features and different classification models. *Biomed. Signal Proceedings and Control*, *18*, 254–262.

Xu, H., Das, A., Saenko, K. (2017). R-C3D: region convolutional 3d network for temporal activity detection. CoRR arXiv:1703.07814.

Yam, C., Nixon, M.S., Carter, J.N. (2004). Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, *37*(5), 1057–1072.

Yan, C.C., Xie, H., Liu, S., Yin, J., Zhang, Y., Dai, Q. (2017a). Effective uyghur language text detection in complex background images for traffic prompt identification. IEEE Trans. Intelligent Transportation Systems.

Yan, C.C., Xie, H., Yang, D., Yin, J., Zhang, Y., Dai, Q. (2017b). Supervised hash coding with deep neural network for environment perception of intelligent vehicles. IEEE Trans. Intelligent Transportation Systems.

Yan, C.C., Zhang, Y., Xu, J., Dai, F., Li, L., Dai, Q., Wu, F. (2014). A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. *IEEE Signal Process. Lett*, *21*(5), 573–576.

Yan, C.C., Zhang, Y., Xu, J., Dai, F., Zhang, J., Dai, Q., Wu, F. (2014). Efficient parallel framework for HEVC motion estimation on many-core processors. *IEEE Transactions Circuits Syst. Video Techn*, *24*(12), 2077–2089.

Yu, S., Tan, D., Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proceedings of international conference on pattern recognition, vol. 4, pp. 441–444. IEEE*.

Yuan, X., Lai, W., Mei, T., Hua, X., Wu, X., Li, S. (2006). Automatic video genre categorization using hierarchical svm. In *Proceedings of international conference on image processing, pp. 2905–2908. IEEE*.

Zha, Z., Mei, T., Wang, Z., Hua, X. (2007). Building a comprehensive ontology to refine video concept detection. In *Proceedings of the international workshop on multimedia information retrieval, pp. 227–236. ACM*.

Zhang, C., Liu, W., Ma, H., Fu, H. (2016). Siamese neural network based gait recognition for human identification. In *IEEE international conference on acoustics, speech and signal processing, pp. 2832–2836*.

Zhang, D., & Shah, M. (2015). Human pose estimation in videos. In *Proceedings of IEEE international conference on computer vision, pp. 2012–2020*.

Zhang, L., Xiang, T., Gong, S. (2016). Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1239–1248*.

Zolfaghari, M., Oliveira, G.L., Sedaghat, N., Brox, T. (2017). Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. CoRR arXiv:1704.00616.