Expert Systems With Applications 179 (2021) 115057

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multi-view gait recognition system using spatio-temporal features and deep learning

Saba Gul [a,*], Muhammad Imran Malik [a], Gul Muhammad Khan [b], Faisal Shafait [a]

[a] School of Electrical Engineering & Computer Science, National University Of Sciences and Technology, Islamabad, Pakistan
[b] National Center of AI, University of Engineering and Technology, Peshawar, Pakistan

ARTICLE INFO

Keywords:
3D convolutional deep neural network (3D CNN)
Gait bio-metric
Gait energy image
Person identification
Optimization

ABSTRACT

Systems based on physiological biometrics are ubiquitous but requires subject cooperation or high resolution to capture. Gait recognition is a great avenue for identification and authentication due to uniqueness of individual stride in an un-intrusive manner. Machine vision systems have been designed to capture the uniqueness of stride of a specific person but factors such as change in speed of stride, view point, clothes and carrying accessories make gait recognition challenging and open to innovation. Our proposed approach attempts to tackle these problems by capturing the spatio-temporal features of a gait sequence by training a 3D convolutional deep neural network (3D CNN). The proposed 3D CNN architecture tackles gait identification by employing holistic approach in the form of gait energy images (GEI) which is a condensed representation capturing the shape and motion characteristics of the the human gait. The network was evaluated on two of the largest publicly available datasets with substantial gender and age diversity; OULP and CASIA-B. Optimization strategies were explored to tune the hyper-parmeters and improve the performance of the 3D CNN network. The optimized 3D CNN and the GEI were effectively able to capture the unique characteristics of the gait cycle of an individual irrespective of the challenging covariates. State of the art results achieved on the multi-views and multiple carrying conditions of the subjects belonging to CASIA-B dataset demonstrating the efficacy of our proposed algorithm.

1. Introduction

In recent years, various measures have been taken to curb terrorist activities by designing automatic screening and surveillance systems. Many screening systems involve authentication or identification of people through their biometrics. Biometric classification comes under two categories: physiological and behavioural. Physiological biometrics refer to the physical attributes of a person (such as face, fingerprint, hand geometry, Iris, and DNA) while behavioral biometrics encapsulate human actions and mannerisms (handwriting, signature, keystroke, voice, and gait).

Systems based on physiological biometrics are commonly deployed for security clearance, but they often require subject cooperation and sometimes even a contact (for example, in fingerprint recognition) (Thapar, Nigam, Aggarwal, & Agarwal, 2018). In many cases where physiological biometrics (such as face recognition) was applied for non cooperating subjects (like persons walking on streets), monitoring through facial recognition is not a feasible option anymore since it requires a controlled environment for accurate and efficient authentication. Efficacious surveillance and authentication systems usually require biometrics that are non-invasive and does not require subject cooperation.

A person's manner of walking is unique and can be used for identification. Gait is an individual's behavioral attribute and it can't be impersonated or changed for a longer span of time. Furthermore, its low susceptibility to noise and ease of access makes gait an adequate biometric to be employed in video surveillance (Thapar et al., 2018).

To date, two methodologies are employed for gait recognition: model-based and holistic or model free approach. Model based approach is based on extracting dynamic information of the human anatomy from the images and tracking changes in these structures over time thus making it robust to noise and occlusion (Yam, Nixon, & Carter, 2004). While Holistic approach takes into account the entire motion pattern of the human body. In comparison to model based approach, it is computationally efficient and can handle low resolution images thus making it suitable for outdoor surveillance.

* Corresponding author.
E-mail addresses: sabagul@uetpeshawar.edu.pk (S. Gul), malik.imran@seecs.edu.pk (M.I. Malik), gk502@uetpeshawar.edu.pk (G.M. Khan), faisal.shafait@seecs.edu.pk (F. Shafait).

https://doi.org/10.1016/j.eswa.2021.115057
Received 18 January 2021; Received in revised form 10 March 2021; Accepted 15 April 2021
Available online 21 April 2021

During the ongoing pandemic of COVID-19, the governments all around the world are taking steps to identify the virus (Altan & Karasu, 2020) and contain the outbreak. To control the spread, it has been made mandatory to wear masks which makes it difficult to identify people through the ubiquitous networks of CCTV cameras already in place. In such a scenario, gait can be considered as a great avenue to identify people in a non-invasive and discreet manner with the already established network of CCTV cameras (Bashir, Xiang, & Gong, 2010).

In this article, we have employed the holistic approach by transforming human silhouettes into gait energy images (GEI) for identification which in contrast to the model-based approach has lower computational cost and is not limited to indoor surveillance making it an ideal approach to be incorporated with the existing CCTV cameras. However, performance of gait recognition algorithms in unconstrained open environments can be degraded with covariates such as multiple view points, variation in illumination, clothing and carrying conditions which makes person identification challenging and open to innovation (Zhang, Liu, Ma, & Fu, 2016). To tackle these challenges, we attempt to capture the spatial features such as body shape and temporal characteristics of walking pattern (spatio-temporal features) (Suresha, Kuppa, & Raghukumar, 2020; Tran, Bourdev, Fergus, Torresani, & Paluri, 2015) through 3D CNN and GEI which captures the information of the entire gait cycle in a condense manner with low susceptibility to noise and occlusion. An efficacious 3D CNN based compact architecture tuned using optimization strategies is proposed for multi-view gait recognition on two of the most challenging datasets: OULP and CASIA-B.

2. Related Work

A novel method preserving spatio-temporal features known as gait energy image (GEI) was introduced by (Han & Bhanu, 2006) in an attempt to identify subjects through their gait by characterizing their walking pattern. A temporal template was created using motion energy image in-order to capture the motion in the sequence of images and a motion history image which captures the recency of the motion taking place to identify human movement by matching them against a set of known movement models (Bobick & Davis, 2001).

The consequences of observing a video over a specific length or a number of gait sequences and its impact on gait recognition is explored to determine a discriminative GEI for identification of a gait (Martin-Felez, Orteils, & Mollineda, 2012). A novel approach to extract features has been described which requires replacing the GEI with GHEI by computing HOG which in addition to preserving information at the boundary, takes into account the inside of the silhouette. The demonstration of the system shows superior results in contrast to its counterpart, GEI (Hofmann & Rigoll, 2012).

The effects of varying co-variate factors within some of the widely known gait datasets are linked to real life non-cooperative subject's behavior for selection of more robust features (Bashir et al., 2010). This scenario when tested on current systems resulted in a drastic degradation in performance. A network is proposed in which features are computed across galleries followed by Adaptive Component and Discriminant Analysis (ACDA) for gait recognition; thus, outperforming its contemporary state of the art systems.

Person identification using reference descriptor (RD) is introduced which instead of direct identification translates the images to similarity measures between the probe and the gallery image (An, Kafai, Yang, & Bhanu, 2016). Regularized canonical correlation analysis (RCCA) is used to generate a subspace in which the correlation between similar subjects is increased. Similarities between probe and gallery images are computed by their re-projection on the RCCA subspace and the RDs. After comparing the RDs of the two sets, a re-ranking step follows to enhance the results and the network is observed to surpass its contemporary frameworks.

Change in walking speed of a person is one of the challenges to accurate identification of a person through gait. Instead of using classification techniques based on distance which is sensitive to starting point of gait cycle; frame-based approach is employed in (Kovac, Štruc, & Peer, 2019). In addition, features are filtered using wavelet approximation resulting in speed invariant discriminative features which does not require training.

The work of Castro, Marín-Jiménez, and Medina-Carnicer (2014) diverges from the conventional method of using binary silhouettes for gait recognition and goes for employing motion descriptors to delineate unique spatial configurations of the descriptors around the person of interest. Fisher vector encoding combines the local motion features extracted across multiple views to obtain high level motion descriptor; the process is termed as Pyramidal Fisher Motion (PFM).

Convolutional neural networks (CNN) show promising results when it comes to learning features in images and is thus extended to the video classification domain (Karpathy et al., 2014; Suresha et al., 2020). Various fusion strategies are evaluated by extending the connectivity of CNN across time domain to explore their input in reducing the processing time followed by evaluation of multi-resolution CNNs. It is observed that high frequency details help in improving the accuracy while lower resolution provides better run-time. Thus, a mixed resolution architecture with low resolution context and high-resolution fovea stream was introduced which performed better by improving the runtime without a detrimental effect on accuracy.

GEI are used in most gait identification architecture which captures the spatial characteristics well, but loses some temporal information. For capturing the temporal characteristics, LSTM networks have been widely employed (Altan, Karasu, & Zio, 2021) which can also be extended to capturing the temporal features of a gait cycle. The shortcoming of GEI images to effectively capture temporal information is met by adding LSTM to the pipeline of CNN, which improves performance across view variations (Feng, Li, & Luo, 2016). The CNN are used to extract the heat maps of individual frames which are invariant to changes in clothing or carrying conditions followed by LSTM to capture the temporal gait features which are invariant across two views. The model also leverages unlabeled data to improve performance of gait recognition.

Acquiring a complete gait cycle in practical scenarios for gait analysis is difficult. Since, in real life some of the frames might be subject to occlusions. To cater for this problem, a network of auto-encoders is trained to transform the incomplete GEI to a complete GEI ensuring an improvement in the identification of persons (Babaee, Li, & Rigoll, 2019). As the number of frames increases, the restored GEI get closer to the ground truth GEI with an improved identification performance. Spatio-temporal features such as local directional pattern (LDP) and optical flow motion are used in conjunction with CNN to achieve a robust system to differentiate between the silhouettes of normal and abnormal gaits (Uddin, Khaksar, & Torresen, 2017).

The challenge of large view difference is investigated by (Wolf, Babaee, & Rigoll, 2016). A deep 3D convolution neural network using grey scale images and optical flow provides color in-variance and is able to generalize gait features across multiple views.The idea of Siamese neural network is introduced by (Zhang et al., 2016) which leverages the Siamese networks potential to extract robust features by using contrastive loss function to minimize the loss of a gait belonging to the same person and set the loss to a large value for gaits associated to different persons.

A state of the art network has been proposed by (Thapar et al., 2018) which utilizes the power of deep CNN to determine the angle of motion of the subject and makes use of different CNNs trained on unique angles to identify the subject. Thus, a robust system for gait identification is introduced across multiple views. In contrast to network introduced by (Wolf et al., 2016), it achieves better/comparable results and is more efficient when it comes to computational cost and time. By introducing stereo images or small overlapping frames, the performance of the network is boosted. Zheng et al. (Liu, Zhang, Wu, & Wang, 2015) attempted to enhance re-identification of humans and make it robust by

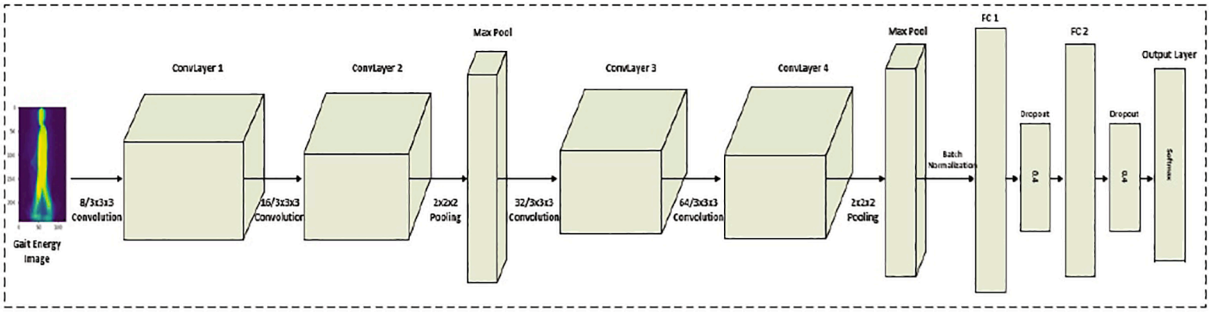


Fig. 1. 3D Convolutional Neural Network for Inter-class Subject Identification.

fusion of information regarding appearance and gait. The feature descriptor comprises of HSV histogram and Gabor feature for appearance and GEI as a gait biometric. The CASIA dataset is used for evaluation which is modified for closed set, open set and cross condition person re-identification. By incorporating gait, in addition to appearance biometric, limitations of appearance-based methods to re-identify persons are overcome. Depth information of images is used to improve the robustness of person re-identification (ReID) using 3D data of face concurrently with skeletal frames of the subjects (Pala, Seidenari, Berretti, & Del Bimbo, 2019). It was observed that when the person of interest is far from the camera, the skeletal frame provides more distinctive information in comparison to when the person is close to the camera. This is because due to close vicinity, some of the skeletal features might be missing and due to clearer facial features, the 3D information of a subject face plays a vital role in person ReID. Thus, by using both the feature cumulatively, state of the art results were achieved. A deep multi-task architecture of CNN is introduced by Marin-Jimenez, Castro, Guil, De La Torre, and Medina-Carnicer (2018). It was observed that by training the network jointly for multiple gait tasks such as age, gender, verification in addition to identification, the training process converges faster with visibly better results. CNN model is introduced in (Castro, Guil, de la Blanca, & Pérez, 2017) to circumvent through the manual process of feature extraction. It makes use of low-level motion features such as optical flow on low resolution images to extract gait signatures to perform identification of gender, subject with different walking style and over elapsed time. The architecture is able to produce state of the art results in comparison to (Castro et al., 2014) with relatively good computational speed. Pose based gait descriptors are extracted from the moving areas around the human joints instead of processing the entire silhouette. It was observed that by considering the moving regions around the joints instead of just processing the entire silhouette through the CNN, an improvement in gait recognition was observed (Sokolova & Konushin, 2019). 3D convolution is explored in (Tran et al., 2015) for action recognition, action similarity labelling; and scene and object identification by varying the network architecture and parameters to determine the best setting. It was observed that C3D introduced performed better than its contemporaries on different video analysis benchmarks.

3. Methodology:

3.1. Prediction model framework

To capture the spatial and temporal features in the gait of an individual, we explored several architectures of CNN. 2D CNN is already well known for its capability to analyze and learn spatial characteristics of an image (Karpathy et al., 2014; Zhang, Yu, & He, 2018). When 2D convolution is applied over successive frames, the temporal information across them is lost. In-order, to explore and learn the temporal dimension as well, we employed a 3D CNN due to its ability to retain time variant information in the GEI (Jia et al., 2014; Thapar et al., 2018; Tran et al., 2015).

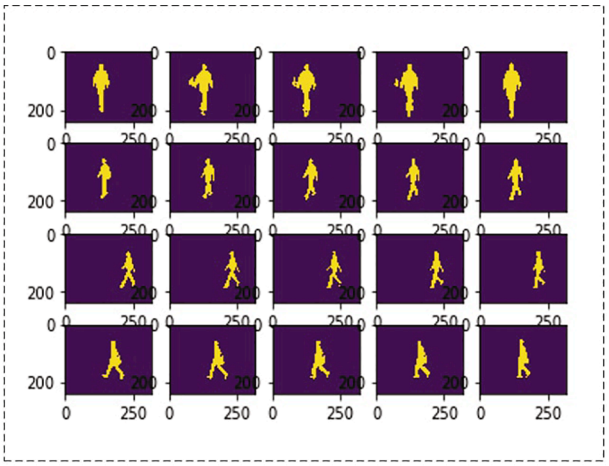


Fig. 2. Raw samples of silhouettes from CASIA-B dataset.

The proposed network comprises of a 3D convolutional neural network inspired from 3D CNN (Tran et al., 2015) but with a compact architecture and lesser parameters to tune by incorporating optimization strategies. The network consists of two sets of two convolutional layers with each set followed by a pooling layer to learn the spatio-temporal features. Batch normalization is applied on the outputs of the convolutional layer before feeding them into the MLP architecture. The last pooling layer is followed by two fully connected layers of 2048 and 512 nodes having dropout and a softmax layer at the end for inter-class classification. Sparse categorical cross entropy and accuracy are used as evaluation metrics (see Fig. 1).

3.2. Dataset specification

The proposed 3D CNN architecture for gait identification is evaluated on two widely used benchmarks in the community for holistic approach: CASIA-B (Yu, Tan, & Tan, 2006) and OULP (Iwama, Okumura, Makihara, & Yagi, 2012) datasets.

CASIA-B is a large multi-view gait database with 102, 828 silhouettes of 124 subjects across 11 different views varying from 0°to 180°. It is the largest dataset with respect to co-variates having gender diversity between men and women (3 : 1) and silhouettes captured across 11

Table 1
Data-set Specifications CASIA-B.

CASIA-B Database Details	
Indoor/Outdoor	Indoor
No. of subjects	124
No. of carrying/walking conditions	3
No.of viewpoints	11

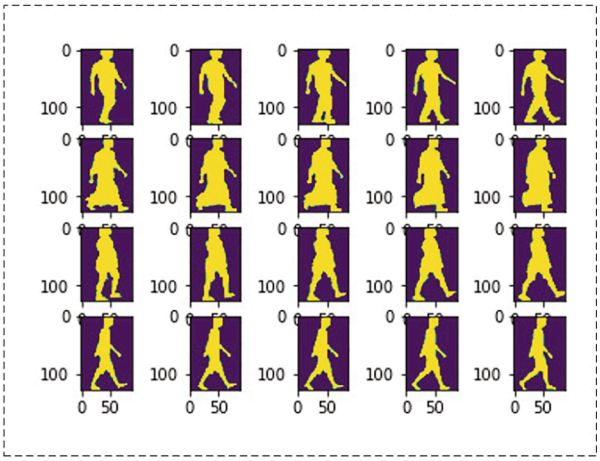


Fig. 3. Samples of size normalized human silhouette from OULP dataset.

Table 2
Data-set Specifications OULP.

OU-ISIR database, Large population dataset Details	
Indoor/Outdoor	Indoor
No. of subjects	4,007(v1), 4,016(v2)
Age range	1–94 years old
No. of carrying/walking conditions	1
No.of viewpoints	4 (55,65,75,85)

different views with 3 different carrying conditions based on clothes and accessories. Fig. 2 depicts the silhouettes from CASIA-B across different angles and walking/carrying conditions with Table 1 depicting the details of this dataset.

The OULP is the largest gait database (w.r.t. number of subjects) with a huge diversity when it comes to gender and age. Moreover, the silhouette quality per subject is relatively high and have undergone a manual visual check. OULP comprises of silhouettes of 4016 subjects belonging to an age group of 1 to 94 years, thus covering a whole generation recorded across 4 different viewpoints. Fig. 3 depicts the silhouettes from OULP dataset with diverse set of clothing and having gender balance ratio close to 1, across different angles with Table 2 having the detailed description of the dataset.

3.3. Preprocessing

The silhouettes of CASIA-B are size normalized to 2 : 1 after object detection with empty frames being discarded. In Fig. 4, size normalization of silhouettes is followed by temporal normalization resulting in gait energy images (GEI) as described by Eq. 1.

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

where N is the total number of frames available in a gait sequence, Ft(x, y) is the t-th frame; x, y are 2-D image coordinates and G(x,y) is the final gait energy image.

OULP dataset comes with size normalized silhouettes of dimensions 88 by 128 which are converted to GEI. Conversion to GEI helps in reducing the appearance of any deformities in the silhouettes due to occlusion, illumination or any noise in the silhouettes.

In CASIA-B, we randomly sample the data and split it into a standard ratio of 70 : 30. After randomly sampling OULP, the generated GEI's are split according to pre-specified gallery and probe set for training and testing, respectively, provided by (Iwama et al., 2012). The OULP-C1V2-All_Gallery is used for training the network while the sequences specified by OULP-C1V2-All_Probe are used for testing.

3.4. Hyperparameter optimization

The 3D CNN architecture used to identify person of interest is tuned to improve its performance by taking into account various optimization strategies (Bergstra, Bardenet, Bengio, & Kégl, 2011; Bergstra, Yamins, & Cox, 2013; Snoek, Larochelle, & Adams, 2012; Chicco, 2017). In-order to find a set of hyperparameters that yield this improvement, Eq. 2 is employed;

$$x^* = argmin_{x\in X} f(x) \tag{2}$$

where f(x) represents the score of objective function that needs to be minimized when evaluated on the validation set; x* is the group of hyperparameters that yields the lowest score, and x represents any value from the X domain. In simple terms, we want to find the model hyperparameters that yield the best score on the validation set metric as described in Fig. 5. Optimization strategies such as grid search, random search and Bayesian optimization are widely used in machine vision and were explored to find optimal set of hyper-parameters(Bergstra et al., 2011; Bergstra et al., 2013).

Bayesian optimization was selected as it was computationally efficient and due to its superior performance on complex, noisy or expensive objective function as shown in Fig. 6. Moreover, unlike the aforementioned optimization strategies, Bayesian optimization is a directed search method as it speeds up the search for optimal hyperparameters by taking into account the performance on previously evaluated hyper-parameters.

In bayesean optimization, we employ Expected Improvement (EI) as an acquisition function which specifies the criteria to select the set of hyperparameters from the surrogate model while Gaussian process model is used as surrogate model for tuning.

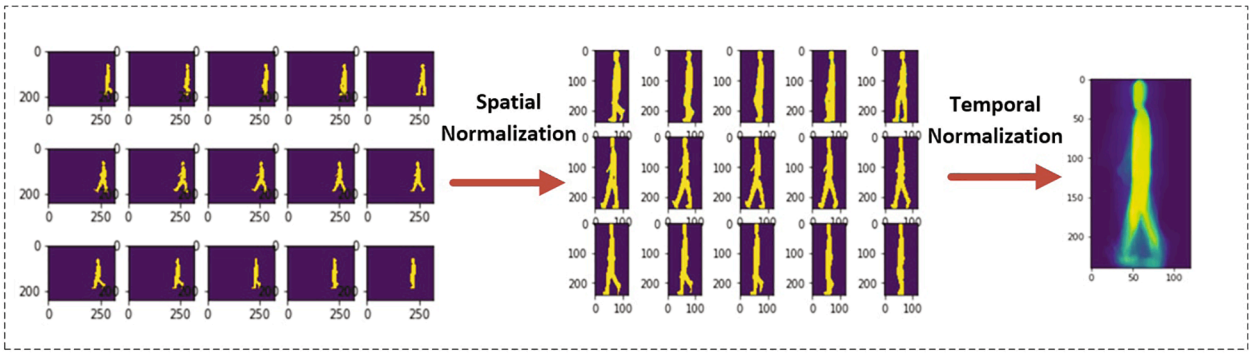


Fig. 4. Preprocessing of Silhouettes From CASIA-B.

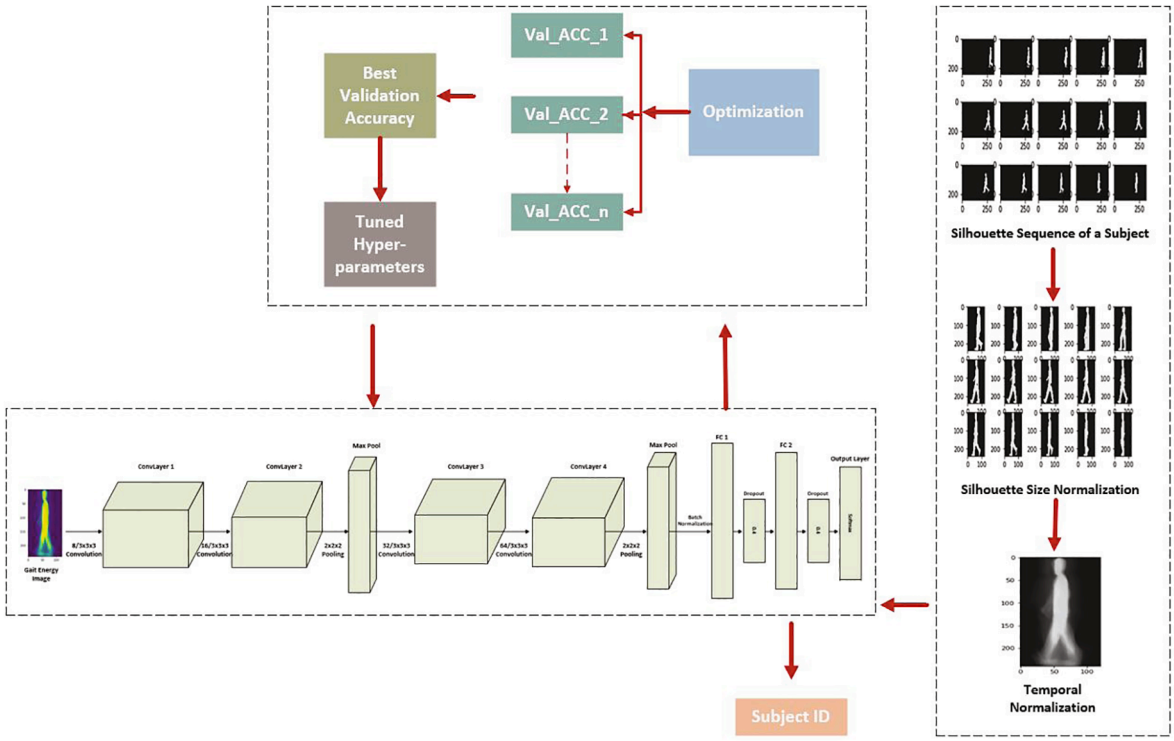


Fig. 5. Tuning Deep Neural Network Using Optimization Strategies.

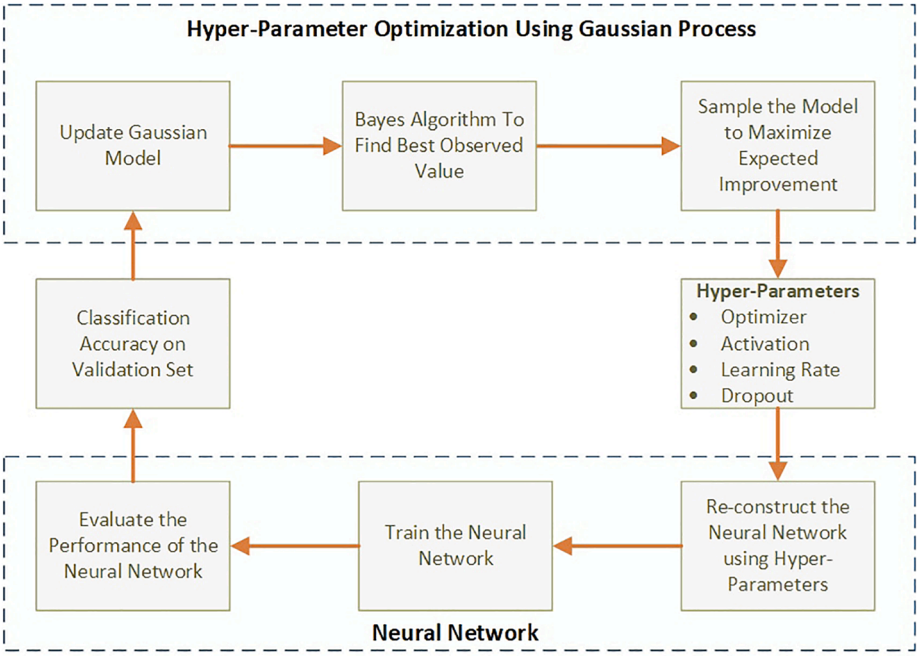


Fig. 6. Hyper-parameter tuning using bayesian optimization.

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) p(y|x) dy \tag{3}$$

where, p(y|x) is the surrogate model, y is the score of the true objective function and x is the hyper-parameter, y* is the minimum score of true objective function observed so far, y is the updated new score.

Optimizer, activation, learning rate, and dropout rate are tuned to get an optimal architecture for gait identification.

4. Results and analysis

The network parameters are tuned using Bayesian optimization to enhance the performance of the model to identify the person of interest. The dataset is fed to the framework after conversion of the silhouettes into GEI's and based on performance of the model on the validation split, optimal hyper-parameters are selected. We start with tuning for the best optimizer and activation followed by learning rate and dropout. The tuned hyper-parameters for CASIA-B dataset are described in

Table 3
Selection of Optimizer and Activation using Bayesian optimization on CASIA-B dataset.

Optimizer	Activation	Accuracy
SGD	*relu*	*0.9808371*
SGD	softsign	0.98008066
Adamax	tanh	0.9783157
Adagrad	relu	0.9778114
SGD	tanh	0.9778114
Adamax	softsign	0.9684821
Adamax	relu	0.9667171
Adadelta	relu	0.96520424

Table 4
Optimal learning rate and dropout selection using Bayesian optimization on CASIA-B dataset.

Learning rate	Dropout	Accuracy
0.03	*0.4*	*0.9760464*
0.013885	0.2	0.9732728
0.00665	0.2	0.9730207
0.01271	0.4	0.9730207
0.009799	0.2	0.9725164
0.005592	0.3	0.9720121
0.011722	0.3	0.97176
0.003004	0.1	0.9715078

Table 3 and Table 4. Using EI as a performance metric of bayesean optimization, the optimizer and activation function are tuned followed by selection of appropriate dropout and learning rate.

After tuning the hyper-parameters, the network is then reconstructed and trained with these optimal hyper-parameters and we observe a boost in its performance. Fig. 7 describe the training and validation performance of the network with Rank-1 accuracy and Rank-5 accuracy of 98.34% and 99.8865% respectively.

The network is then re-tuned for the OULP dataset with the performance gauged on the validation set. The result of Bayesian optimization is presented in Table 5 and Table 6. We started out by finding the optimal optimizer and activation function in the search space. After modifying the network using the optimal activation function and optimizer in Table 5, the network was re-tuned to find the learning rate and dropout to boost the performance of the network.

The framework is then modified by assigning these optimal hyper-parameters and an improvement is observed in the identification performance. Fig. 8 describes the training and validation performance on the OULP dataset with Rank-1 and Rank-5 accuracy of 93.1872% and 98.5054% respectively. The final network parameters used to train the 3D CNN are described in Table 7.

On comparison of our framework with state of the art architectures described in Table 8, we observe that our network surpasses it in case of CASIA-B and achieve appreciable results on OULP by using the same framework. The deterioration in performance on the OULP dataset can be attributed to lesser number of silhouettes per subject which results in fewer GEI's per subject.

The Correct Classification Rate (CCR%) is performed to evaluate the performance of our proposed model on the 11 unique view-points of the CASIA-B dataset.

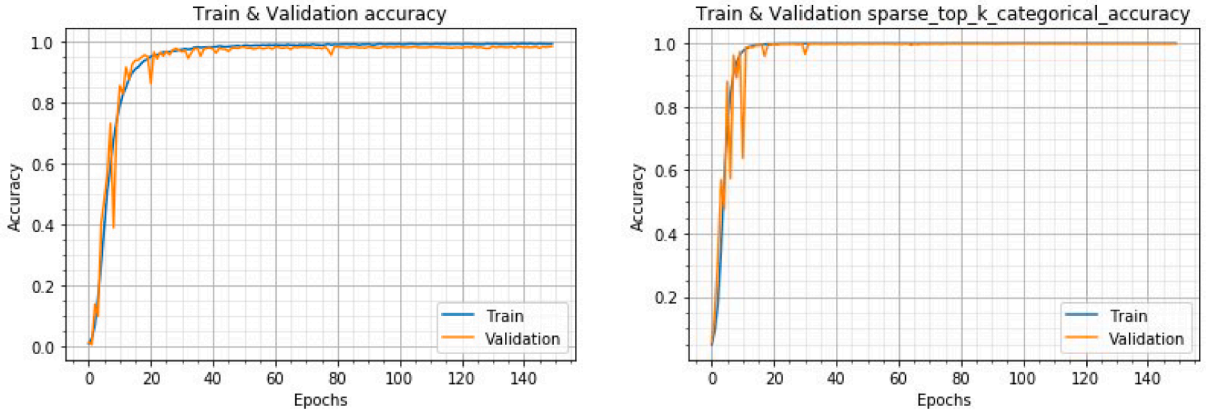
$$CCR = \frac{TP + TN}{TP + FP + FN + TN}(\%) \tag{4}$$

Table 5
Selection of Optimizer and Activation using Bayesian optimization on OULP dataset.

Optimizer	Activation	Accuracy
SGD	*relu*	*0.920044*
SGD	tanh	0.9041812
SGD	linear	0.8928113
SGD	softplus	0.88355035
Adamax	linear	0.87071335
Adamax	softsign	0.8660371
Adadelta	relu	0.8461397
SGD	softsign	0.84082156

Table 6
Optimal learning rate and dropout selection using Bayesian optimization on OULP dataset.

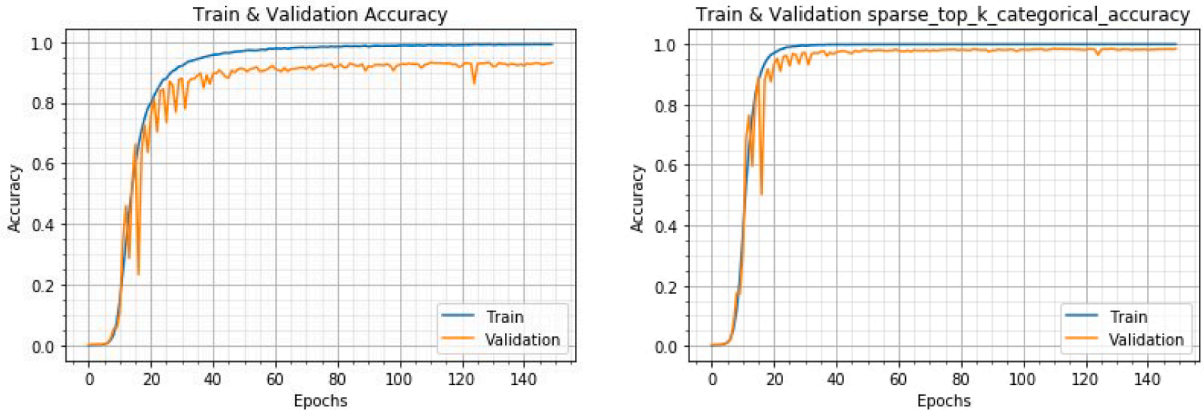
Learning rate	Dropout	Accuracy
0.01	*0.4*	*0.91335046*
0.01	0.2	0.90858245
0.01	0.33	0.908399057
0.01	0.29	0.90748215
0.01	0.32	0.9054649
0.0042694	0.25	0.8936365
0.0029120	0.29	0.8932698
0.01	0.22	0.8926279



(a) Rank 1 accuracy of the network for CASIA-B

(b) Rank 5 accuracy of the network for CASIA-B

Fig. 7. Network training results on CASIA-B dataset.



(a) Rank 1 accuracy of the network for OULP

(b) Rank 5 accuracy of the network for OULP

Fig. 8. Network training results on OULP dataset.

Table 7
Data-set Specifications CASIA-B.

Details of person identification model parameters	
Batch Size	16
Epochs	130 (Casia-B), 150 (OULP)
Optimizer	SGD
Learning rate	0.01
Drop-out rate	0.4

Table 8
Comparison of our framework with state of art models.

Model	Dataset	Rank-1 Accuracy	Rank-5 Accuracy
3D CNN (Thapar et al., 2018)	CASIA-B	97.65%	N/A
3D CNN (Ours)	CASIA-B	98.34%	99.8865%
Siamese NN (Zhang et al., 2016)	OULP	96.02%	98.31%
3D CNN (Ours)	OULP	93.1872%	98.5054%

Table 9
Comparison of quality performance in CCR(%) of our framework on CASIA-B with previous architectures for different view-points.

Angle	(Wolf et al., 2016)	(Thapar et al., 2018)	Stereo images (Thapar et al., 2018)	3D CNN (Ours)
0°	96.30%	98.33%	**98.67%**	97.52%
18°	98.20%	99.17%	**99.55%**	99.17%
36°	98.50%	99.17%	**100%**	99.45%
54°	95.40%	96.67%	**99.55%**	98.89%
72°	94.30%	97.92%	97.78%	**99.45%**
90°	**99.90%**	97.08%	97.79%	99.17%
108°	98.60%	97.91%	**98.67%**	96.48%
126°	97.00%	97.08%	96.90%	**99.17%**
144°	97.40%	96.25%	96.38%	**99.17%**
162°	**99.20%**	96.67%	96.10%	98.61%
180°	96.10%	97.08%	97.69%	**99.44%**

Where TP, FP, FN and TN represent the true positive, false positive, false negative and true negative values of classes respectively.

Table 9 shows a comparison of our model with previous state of the art architectures. On caparison with (Wolf et al., 2016) which makes use of binary silhouettes along with optical flow, our network outperforms it by effectively incorporating the gait cycle information into a GEI. In contrast with VGR-Net (Thapar et al., 2018) which makes use of human silhouettes; our network shows a better performance on most of the view-points despite its compact architecture which can be attributed to effective spatio-temporal feature extraction and learning. It can be observed that VGR-net (Thapar et al., 2018) shows a boost in performance but at an additional expense by using stereo images along with small overlapping clips.

To validate our results further on the gait identification datasets, evaluation metrics of precision, recall and F1 are employed as described by Eq. (5)–(7) respectively.

$$F1 - Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{5}$$

where;

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

Table 10 describes the weighted, micro and macro precision, recall and F1 score respectively to evaluate the performance of multi-class classification problem and further elaborate the results achieved by

Table 10
Evaluation of performance of 3D CNN on CASIA-B.

Micro Precision	0.98071	Macro Precision	0.98013	Weighted Precision	0.98147
Micro Recall	0.98071	Macro Recall	0.98042	Weighted Recall	0.98071
Micro F1-Score	0.98071	Macro F1-Score	0.97990	Weighted F1-Score	0.98072

Table 11
Evaluation of performance of 3D CNN on OULP.

Micro Precision	0.93187	Macro Precision	0.93628	Weighted Precision	0.95167
Micro Recall	0.93187	Macro Recall	0.93537	Weighted Recall	0.93187
Micro F1-Score	0.93187	Macro F1-Score	0.92423	Weighted F1-Score	0.93160

the network in-case of CASIA-B dataset.

The weighted, micro and macro precision, recall and F1 score respectively of OULP dataset are described in Table 11 to elaborate the results achieved.

From Table 9, 10, 11 and 7, it can be drawn that GEI can effectively encapsulate spatio-temporal characteristics of gait cycle by showing resilience to noise, low resolution, illumination and certain occlusions. Thus by fusing GEI and 3D CNN network, we are able to learn the characteristic gait features despite challenges of different covariates in the datasets and outperform the previous architectures at most of the view-points.

5. Conclusion

We have introduced a 3D convolutional neural network to extract robust and discriminative spatio-temporal features for gait recognition to tackle the challenges such as view angle and occlusion due to different clothing and walking conditions. GEI representation is adapted to remove noise and variation in illumination conditions in images while preserving the body shape and walking frequency. The features of the GEI in-addition to the features extracted by the 3D CNN enable the network to understand the gait cycle and identify a person more accurately. Optimization through Bayesian algorithms enables us in narrowing down the search space for an optimal set of hyperparameters in an informed manner, thus boosting the performance of the framework. The results of the experiments on the widely used benchmark datasets demonstrate the effectiveness of our framework with state of the art results achieved on CASIA-B. Nevertheless, overfitting can be seen as potential problem due to small amount of variance and limited number of frames per subject in-case of OULP to compute the GEI. In the future, we wish to explore, genetic optimization algorithms and extend the work to practical environment by working with more challenging real-life scenarios to identify persons through their walking pattern.

CRediT authorship contribution statement

Saba Gul: Data curation, Conceptualization, Methodology, Validation, Writing - original draft. **Muhammad Imran Malik:** Data curation, Supervision, Writing - review & editing. **Gul Muhammad Khan:** Project administration, Writing - review & editing. **Faisal Shafait:** Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Altan, A., & Karasu, S. (2020). Recognition of covid-19 disease from x-ray images by hybrid model consisting of 2d curvelet transform, chaotic salp swarm algorithm and deep learning technique. *Chaos, Solitons and Fractals,*. https://doi.org/10.1016/j.chaos.2020.110071

Altan, A., Karasu, S., & Zio, E. (2021). A new hybrid model for wind speed forecasting combining long short-term memory neural network, decomposition methods and grey wolf optimizer. *Applied Soft Computing*. https://doi.org/10.1016/j.asoc.2020.106996

An, L., Kafai, M., Yang, S., & Bhanu, B. (2016). Person reidentification with reference descriptor. *IEEE Transactions on Circuits and Systems for Video Technology, 26*, 776–787. https://doi.org/10.1109/TCSVT.2015.2416561

Babaee, M., Li, L., & Rigoll, G. (2019). Person identification from partial gait cycle using fully convolutional neural networks. *Neurocomputing, 338*, 116–125. https://doi.org/10.1016/j.neucom.2019.01.091. arXiv:1804.08506

Bashir, K., Xiang, T., & Gong, S. (2010). Gait recognition without subject cooperation. *Pattern Recognition Letters, 31*, 2052–2060. https://doi.org/10.1016/j.patrec.2010.05.027

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011, (pp. 1–9).

Bergstra, J., Yamins, D., & Cox, D.D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In 30th International Conference on Machine Learning, ICML 2013.

Bobick, A. F., & Davis, J. W. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence, doi, 10* (1109/34), Article 910878.

Castro, F. M., Marín-Jiménez, M. J., & Medina-Carnicer, R. (2014). Pyramidal fisher motion for multiview gait recognition. *Proceedings - International Conference on Pattern Recognition*. https://doi.org/10.1109/ICPR.2014.298. arXiv:1403.6950.

Castro, F. M. M.-J., Guil, M. J., de la Blanca, N., & Pérez, N. (2017). Automatic learning of gait signatures for people identification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-59147-6_23. arXiv:1603.01006.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. doi: 10.1186/s13040-017-0155-3.

Feng, Y., Li, Y., & Luo, J. (2016). Learning effective gait features using lstm. In *Proceedings - International Conference on Pattern Recognition*. https://doi.org/10.1109/ICPR.2016.7899654

Han, J., & Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TPAMI.2006.38

Hofmann, M., & Rigoll, G. (2012). Improved gait recognition using gradient histogram energy image. In Proceedings - International Conference on Image Processing, ICIP. doi: 10.1109/ICIP.2012.6467128.

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*. https://doi.org/10.1109/TIFS.2012.2204253

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *In MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*. https://doi.org/10.1145/2647868.2654889. arXiv:1408.5093.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Li, F. F. (2014). Large-scale video classification with convolutional neural networks. In *In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2014.223

Kovac, J., Štruc, V., & Peer, P. (2019). Frame-based classification for cross-speed gait recognition. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-017-5469-0

Liu, Z., Zhang, Z., Wu, Q., & Wang, Y. (2015). Enhancing person re-identification by integrating gait biometric. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2015.05.008

Marin-Jimenez, M. J., Castro, F. M., Guil, N., De La Torre, F., & Medina-Carnicer, R. (2018). Deep multi-task learning for gait-based biometrics. *Proceedings - International Conference on Image Processing, ICIP.*. https://doi.org/10.1109/ICIP.2017.8296252

Martin-Felez, R., Orteils, J., & Mollineda, R. A. (2012). Exploring the effects of video length on gait recognition. In *In Proceedings - International Conference on Pattern Recognition*.

Pala, P., Seidenari, L., Berretti, S., & Del Bimbo, A. (2019). Enhanced skeleton and face 3d data for person re-identification from depth cameras. *Computers and Graphics (Pergamon)*. https://doi.org/10.1016/j.cag.2019.01.003

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *In Advances in Neural Information Processing Systems. <hi rend="tt">arXiv:1206.2944</hi>*.

Sokolova, A., & Konushin, A. (2019). Pose-based deep gait recognition. *IET Biometrics*. https://doi.org/10.1049/iet-bmt.2018.5046. arXiv:1710.06512.

Suresha, M., Kuppa, S., & Raghukumar, D. S. (2020). A study on deep learning spatiotemporal models and feature extraction techniques for video understanding. *International Journal of Multimedia Information Retrieval*. https://doi.org/10.1007/s13735-019-00190-x

Thapar, D., Nigam, A., Aggarwal, D., & Agarwal, P. (2018). Vgr-net: A view invariant gait recognition network. In 2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis, ISBA 2018. doi: 10.1109/ISBA.2018.8311475. arXiv:1710.04803.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2015.510. arXiv:1412.0767.

Uddin, M. Z., Khaksar, W., & Torresen, J. (2017). A robust gait recognition system using spatiotemporal features and deep learning. *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*. https://doi.org/10.1109/MFI.2017.8170422

Wolf, T., Babaee, M., & Rigoll, G. (2016). Multi-view gait recognition using 3d convolutional neural networks. In *Proceedings - International Conference on Image Processing*. ICIP. https://doi.org/10.1109/ICIP.2016.7533144.

Yam, C. Y. Y., Nixon, M. S., & Carter, J. N. (2004). Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2003.09.012

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 - International Conference on Pattern Recognition*. https://doi.org/10.1109/ICPR.2006.67

Zhang, C., Liu, W., Ma, H., & Fu, H. (2016). Siamese neural network based gait recognition for human identification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. https://doi.org/10.1109/ICASSP.2016.7472194

Zhang, W., Yu, X., & He, X. (2018). Learning bidirectional temporal cues for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*. https://doi.org/10.1109/TCSVT.2017.2718188