Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.016574
Article

Tech Science Press

CAMNet: DeepGait Feature Extraction via Maximum Activated Channel Localization

Salisu Muhammed* **and Erbuğ Çelebi**

Computer Engineering Department, Cyprus International University, Nicosia, North Cyprus, Mersin 10, 099010, Turkey
*Corresponding Author: Salisu Muhammed. Email: 21700969@student.ciu.edu.tr
Received: 05 January 2021; Accepted: 05 February 2021

Abstract: As the models with fewer operations help realize the performance of intelligent computing systems, we propose a novel deep network for DeepGait feature extraction with less operation for video sensor-based gait representation without dimension decomposition. The DeepGait has been known to have outperformed the hand-crafted representations, such as the frequency-domain feature (FDF), gait energy image (GEI), and gait flow image (GFI), etc. More explicitly, the channel-activated mapping network (CAMNet) is composed of three progressive triplets of convolution, batch normalization, max-pooling layers, and an external max pooling to capture the Spatio-temporal information of multiple frames in one gait period. We conducted experiments to validate the effectiveness of the proposed novel algorithm in terms of cross-view gait recognition in both cooperative and uncooperative settings using the state-of-the-art OU-ISIR multi-view large population OU-MVLP dataset. The OU-MVLP dataset includes 10307 subjects. As a result, we confirmed that the proposed method significantly outperformed state-of-the-art approaches using the same dataset at the rear angles of 180, 195, 210, and 225, in both cooperative and uncooperative settings for verification scenarios.

Keywords: Feature extraction; gait representation; channel; frame; systems

1 Introduction

The behavioral characteristic of a person's pattern of walking known as gait is unique to individuals due to its reliability in verifying identity. However, concern about privacy has been an issue during the collection of such biometric identifiers. Unlike the traditional and knowledge base identification approaches that required subject cooperation, gait information can be collected from a distance at a low resolution without subject cooperation.

Appearance-based and model-based are the two methods employed for video sensor-based gait recognition. The motion of a human body is the point of interest in appearance-based and it operates on silhouettes. Silhouette extraction, gait period detection, gait representation generation, and recognition are the sequence in the framework of the appearance-based method.

The model-based method which is computational expensive focuses mainly on strides parameters extraction from a subject to describe the pattern and hence requires high-resolution images. As the gait recognition needs to be real-time and effective at low resolution, various algorithms as well as frameworks were proposed to learn rich features using neural networks and deep learning. Some frameworks were proposed because not every neural network architecture can fit into real-time applications. The real-time performance can be achieved with models that have fewer operations. The well-known pre-trained networks such as the AlexNet, GoogleNet are not designed with fewer operations. In essence, studies on gait recognition that utilize a deep learning framework are quite a few since it requires adequate training samples. It is however difficult to collect adequate gait samples for training.

Gait features are extracted from a defined gait period where silhouettes are aggregated together in the item of the gait energy image (GEI) [1], it comprises both static and dynamic parts of human motion information. Approaches exist that focused only on the dynamic features to keep aside the static information such as the clothing and carrying condition. In the gait flow image (GFI) approach, the sequence of binary images of a moving contour is aggregated for a single gait period [2] while the chrono-gait image (CGI) on the contrary aggregates the binary image sequence along the contour of one-fourth of gait period using color encoding strategy. These approaches use contour to represent the walking pattern. Moreover, the dynamic part of the GEI was extracted [3] to form a gait entropy image (GEnI), where information theory is used to highlight the dynamic parts which are of high and low intensities.

A generative approach to robust gait recognition is meant for specific covariates such as the walking speed, view angle, and carrying condition. It is however not optimally guaranteed for gait recognition because of its aim at frame synthetization. It also optimized the accuracy of the generated features, not discrimination capability. In the case of view angle covariate, the probe and gallery view angles are to be known in advance. However, a View Transformation Model VTM for cross-view gait recognition with a quality measure in cooperated to encode the level of the bias was proposed by [4]. The two quality measures were calculated using the pair of the gait features while the posterior probability that both the gait originated from the same subject was calculated using the two quality measures [4].

To extract invariant gait features, [5] proposed a GaitGAN which is based on generative adversarial networks (GAN), and took the GAN model as a regressor to generate invariant gait images that side view images. Their approach differs from the traditional GAN as it has two discriminators while the GAN has only one. The GaitGAN is shown to be promising for practical applications in video surveillance. Moreover, [6] upgraded the GaitGAN to GaitGANv2 where there is no need to define the view angle before generating invariant gait images and also signifies an improvement over GaitGANv1 in such a way that, the former adopts a multi-loss strategy to optimize the network to increase the inter-class distance and to reduce the intra-class distance concurrently. As most of the obtainable models need to estimate the view angle first and can work for only one view pair. [7] Employed one deep model based on auto-encoder for invariant gait extraction. Their model was able to synthesized gait feature progressively using auto-encoders stacked in a multi-layer manner. The unique advantage of their model is that it can extract invariant gait features using only one model.

The autoencoder was also used as a framework by [8] to explicitly disentangle pose and appearance features from RGB imagery. They also proposed the LSTM-based integration of pose features over time produces the gait feature which can feature disentanglement qualitatively. [9] Also used disentangled representation learning to propose a method of gait recognition that considers both identity and covariate features. It avoids failure modes when variations occur due to the covariate overwhelm. They initially coded an input gait pattern to get the disentangled identity and covariate features. [9] Further decrypted the features to simultaneously reconstruct the input gait pattern and canonical version of the same subject with no co-variates in a semi-supervised manner to ensure successful disentanglement. [10] Proposed a

novel network termed GaitSet to study identity information from the set. Their method has immunity to permutation of frames, and can also naturally incorporate frames from different videos captured under different scenarios. On the other hand, a part-based model termed GaitPart was proposed by [11] to enhance the fine-grained learning of the part-level spatial features and on the other hand, the Micro-motion Capture Module (MCM) to the pre-defined parts of the human body.

A pairwise spatial transformer network PSTN as a unified CNN architecture that consists of a pairwise spatial transformer (PST) and subsequent recognition network (RN) was proposed by [12]. When given a similar pair of gait features from a different source and target views, the PST guesses a non-rigid deformation field to record the features in the similar pair into their intermediate view. [13] Extracted the VTM as a frequency domain feature from the silhouette volume of a moving person. Training subset of multiple individuals from the various direction was used to generate the model which transforms the gallery features into the same angle as the input or probe features during the registration stage. Partial least square was used by [14] as a feature selection method for establishing a robust VTM with Principal Component Analysis PCA. The construction of the robust VTM, feature selection, and view transformation projection was learned during registration. This method is capable of identifying a share linear correlated low-rank subspace [14]. A GEI and Truncated Singular Vector Decomposition TSVD was considered during their setup. Similarly, [15] adopted the singular value decomposition technique in creating a VTM based spatial domain GEI. They used the VTM to transform the view angles of both gallery and probe gait data into the same direction to enable the measurement of gait signatures without constraints. [16] Developed a new method referred to as Complete Canonical Correlation Analysis (C3A) to overcome the limitation of PCA and SVD. Singular generalized Eigensystem computation of canonical correlation analysis transformed into stable decomposition problems. They stated that the proposed flexible and effective technique alleviates the burden of high-dimensional matrix computation.

The discriminative approach is contrary to the generic approach is for general covariate where Linear Discriminant Analysis LDA is often applied and might be necessary to also apply PCA when high dimensional features of gait are involved. [17] Used the discriminative projection with List-Wise constraint for learning projection capable of mapping gait representations from various views into a discriminative subspace common to them. A gait individuality image gets rid of the redundant view-dependent part from GEI and reserve the individual identification information. [18] Used a deep learning framework to obtain high-level features for multiple view gait recognition incremental model. [19] Proposed an appearance-based gait recognition method where deep features were extracted using simple CNN that has different activation functions within the same structure. They used 10 layers with 240 $\times$ 240 grayscale input images from the CASIA-A dataset. [20] Proposed a model-based gait recognition where inertia gait data were extracted. They also proposed a hybrid network for a robust gait feature representation that successively abstracted features in both time and space domains using RNN and CNN. 118 subject information was collected using an application installed on a smartphone. [21] Proposed an improved DNN that can overcome the covariate factors in gait recognition where subject cooperation is not needed. Their model was able to learn discriminative features and promising results were obtained. Time-based graph long short term memory was used by [22] to develop a robust deep learning model for 3d MSR action dataset for action recognition. CASIA-B and TUM-GAID datasets were used to evaluate the model for gait recognition. [23] Used deep learning to explain the unique gait pattern of the individual. The classification was based on kinematic data and time-continuous kinetic data for gait pattern recognition. 57 individuals were used free of flower extremity injuries and without gait pathology. [24] Proposed a Spatial-Temporal Gradient Feature STGF to get rid of optical flow feature extraction as it requires big computing resources due to its complex mathematical calculation in a video-related task. Spatial Feature Network SFN was used to extract the spatial features from GEIs by applying them to

CNN whereas, Temporal Feature Network TFN was used to extract the temporal features from silhouette frames. The two features from SFN and TFN were fused to come up with STGF.

A comprehensive study was carried out by [25] using DNN for cross-view gait human gait identification. Cross walking conditions and cross-view scenarios were proposed using three gait databases. Three different network architectures involved deep as well as nonlinear matching were investigated with where and when to start the matching as the major distinction between the network architectures. [26] Proposed a DeepGait representation for view-invariant gait recognition where features were captured by forward propagating the gait images through a pre-trained VGG-D 19 network with 19 layers.4096-dimensional features were obtained at the fully connected layer of the network.256 dimension DeepGait with Joint Bayesian gave the highest performance when PCA is in cooperated. Hence complicated the model structure in terms of computational complexity. Our novel CAMNet makes the DeepGait more suitable for real-time performance as it is being produced with fewer operations unlike the [26] produced using 19 layers and external max pooling operation.

The proposed deep network operates under the appearance-based technique for gait recognition. Besides, what distinguishes it from the previous contribution in the field of deep learning for gait feature extraction is that each silhouette frame in a gait period is being propagated forward through the CAMNet and the maximum activated channel is being extracted from the feature maps at the last ReLu layer. This is because the ReLu activation function converges fast and also to utilize it sparsely advantage of removing all the negative pixel values redundant feature. Max pooling is employed to combine the Spatio-temporal information in one gait period [26]. This process alleviates the need for dimension decomposition in the system. Moreover, the big computing resources needed by the PCA in [26] due to its complex mathematical calculation in a video-related task have been solved.

To summarize, our contributions are as follows:

1. We improved the GEINet by adding a batch normalization layer to speed up the feature extraction on a deep learning framework.
2. We proposed CAMNet, a channel-discriminative localization technique that generates visual explanations of an improved GEINet and also gives room for extracting maximum activated channel within a feature map.
3. The channel-discriminative localization technique proposed alleviates the need for dimension decomposition during post-processing in any deep learning framework.
4. We evaluated the CAMNet on the state-of-the-art gait dataset OU-MVLP and used it to develop a new DeepGait for 14 different angles in two sequences.
5. We developed an algorithm for normalizing silhouette frames of OU-MVLP world's largest public gait dataset.

The rest of the paper is organized as follows. Section 2 introduces Materials and Methods. Section 3 Experimental results. Section 4 presents the discussions. Finally, Section 5 states conclusions with future trends.

2 Materials and Methods

2.1 Multi-view Large Population Dataset

The dataset was downloaded from the Osaka university OU-ISIR Gait database. After signing a release agreement and sent to the database administrator. The password was then released to enable us access to the downloaded dataset.

2.1.1 Capturing System

The data was collected in conjunction with an experience-based long-run exhibition of video-based gait analysis at a science museum. The approved informed consent was obtained from all the subjects in this dataset. The dataset consists of 10,307 subjects (5,114 males and 5,193 females with various ages, ranging from 2 to 87 years) from 14 view angles, ranging 0°–90°, 180°–270°. Gait images of 1,280 × 980 pixels at 25 fps are captured by seven network cameras (Cam1-7) placed at intervals of 15° azimuth angles along a quarter of a circle whose center coincides with the center of the walking course. Its radius is approximately 8 m and its height is approximately 5 m [27]. The subject repeats forward (A to B) and backward (B to A) walking twice of each, 28 gait image sequences (=7 (cameras) × 2 (forward and backward) × 2 (twice)) can be captured per one subject. The view angle of our dataset is defined in Fig. 1 [27].

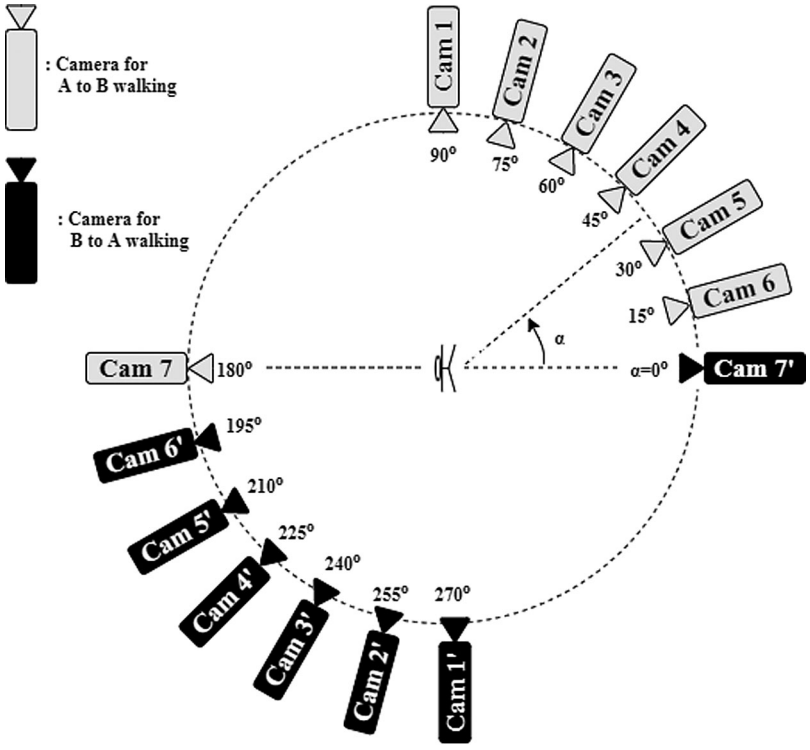

Figure 1: A definition of view angles

2.1.2 The Procedure of GEI Formation

The procedure for the dataset's GEI formation is explained in the following subheadings.

2.1.2.1 Silhouette Sequence Extraction

They extracted the human region using the chromo-key technique by removing the green part according to HSV color space within the gait course. Fig. 2 shows the extracted silhouette.

2.1.2.2 Size Normalization

"Silhouette extracted and regularized by size based on the method used in [28]. First, the top, bottom, and horizontal center of the silhouette regions are obtained for each frame. Second, a moving-average filter is applied to these positions. Third, they regularized the size of the silhouette images such that the height is just 128 pixels according to the average positions, and the aspect ratio of each region is maintained. Finally, we produce an 88 × 128 pixel image in which the average horizontal median corresponds to the horizontal center of the image."

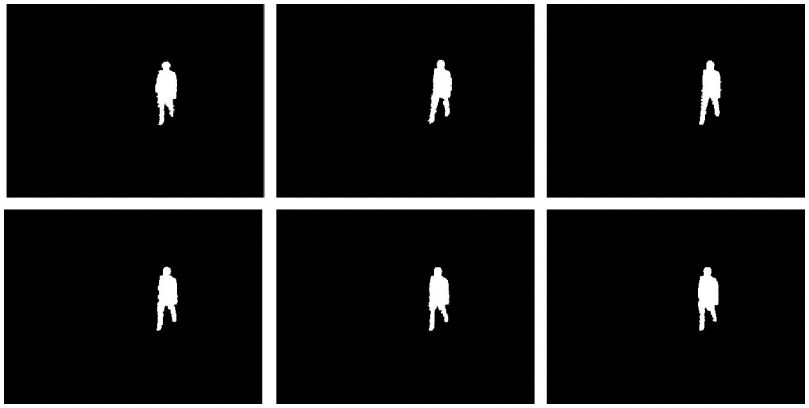

Figure 2: An illustration of silhouette sequence at 45°

2.1.2.3 Gait Period Detection

"The Gait period was detected based on size-normalized silhouette sequences from cam1 by using the method proposed in [28]. They then adopted the normalized autocorrelation (NAC) of the size-normalized silhouette images for the temporal axis and also obtained the gait period as the frameshift corresponding to the second peak of the NAC" [27].

2.1.2.4 GEI Extraction

"The GEI was extracted by averaging the size- normalized silhouette sequences pixel-wise over one gait period" [27]. The nearest gait period to the center of the walking course is used when several gait periods are detected from one walking sequence. A generated GEI for view angle 45° is shown in Fig. 3.

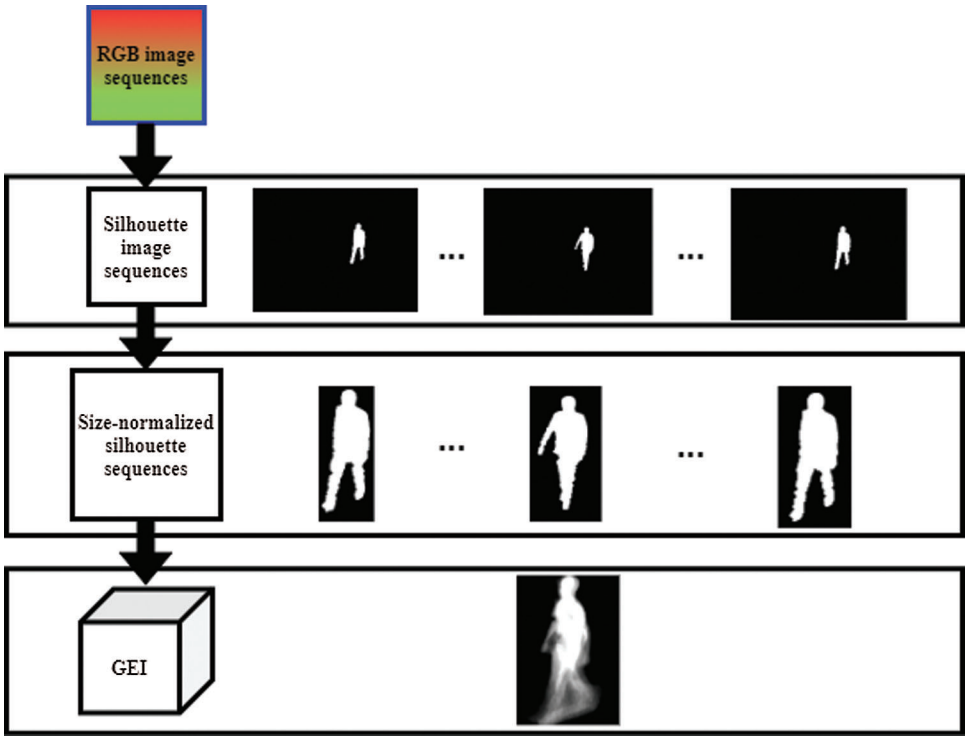

Figure 3: An illustration of GEI extraction flow

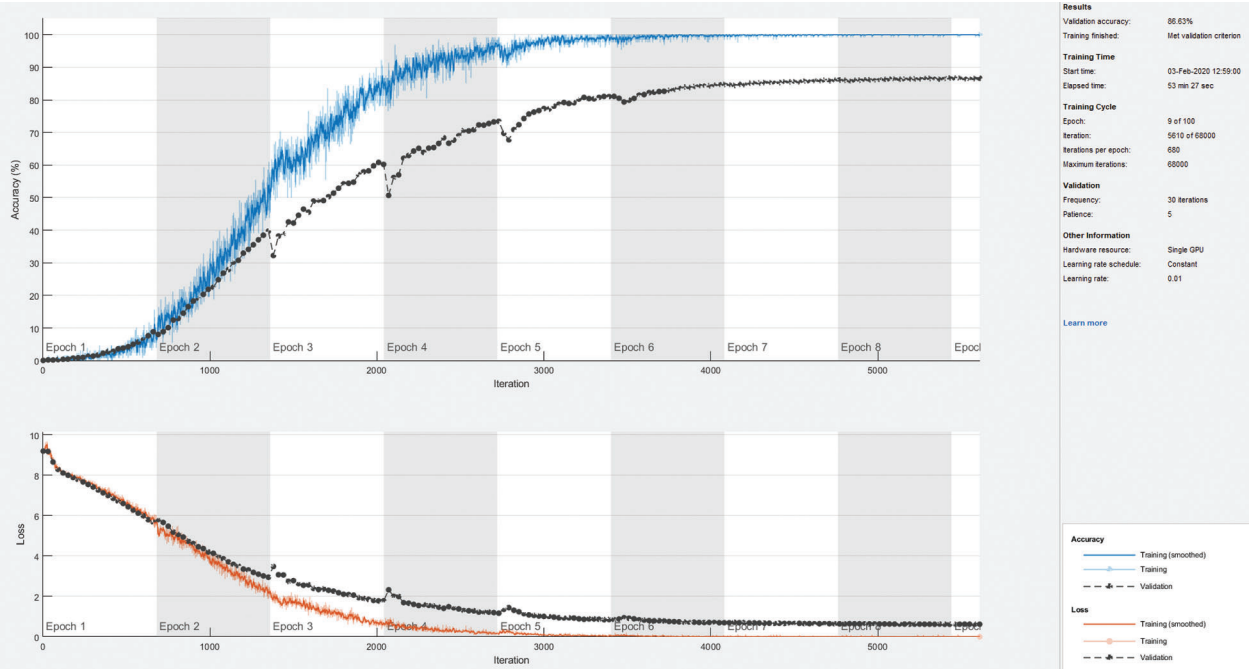

Figure 4: The training progress plot showing the mini-batch loss and accuracy and the validation loss and accuracy

2.2 Input Data

The input to the improved GEINet is grayscale GEI images obtained from the state-of-the-art OU-ISIR multi-view large population dataset as shown in Fig. 5.

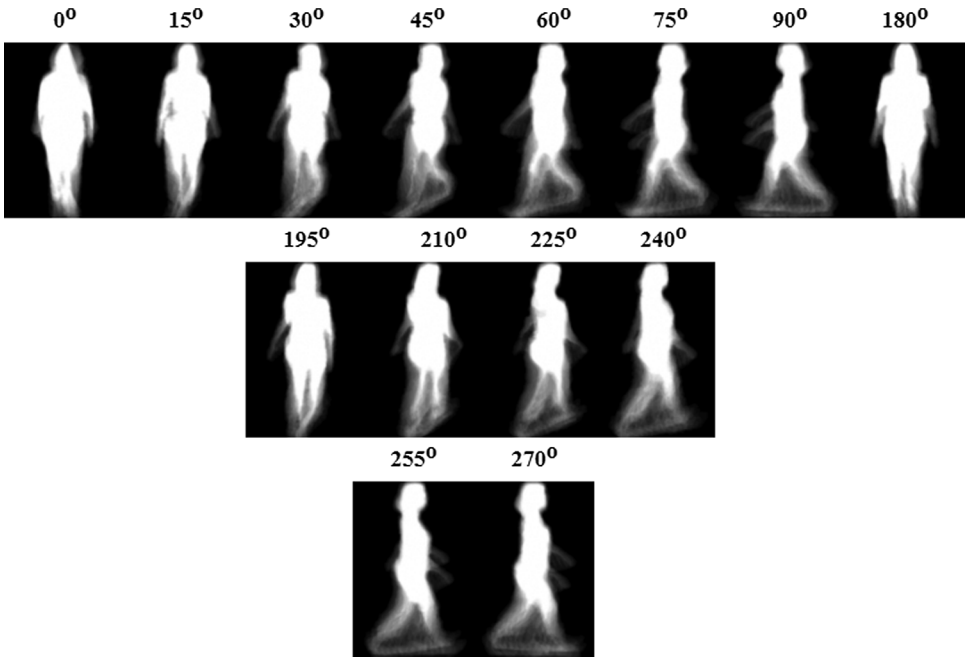

Figure 5: An illustration of OU-MVLP GEIs

2.3 Network Structure

The improved GEINet is structured as a fifteen-layered CNN network, whereby the leading layers are progressive triads of input, convolution, batch normalization, activation, and max-pooling layers as shown in Fig. 6. The variable settings for training and validation of the improved GEINet as well as the outlines for each convolution and pooling layer are illustrated in Tabs. 1 and 2 respectively. The ReLu activation function is used. In the learning phase, a set of similarities to individual training subjects is calculated using the softmax function. Improved GEINet design has a similar structure, as one of the famous network structures, that is, AlexNet used for classification of images [29]. However, improved GEINet is not as deep compared with recent deep neural networks such as AlexNet. This is because AlexNet concentrations on the image classification task under large spatial displacement, whereas improved GEINet concentrations on subtle inter-subject variances within the same action class, that is, gait, using spatially well-aligned silhouettes.

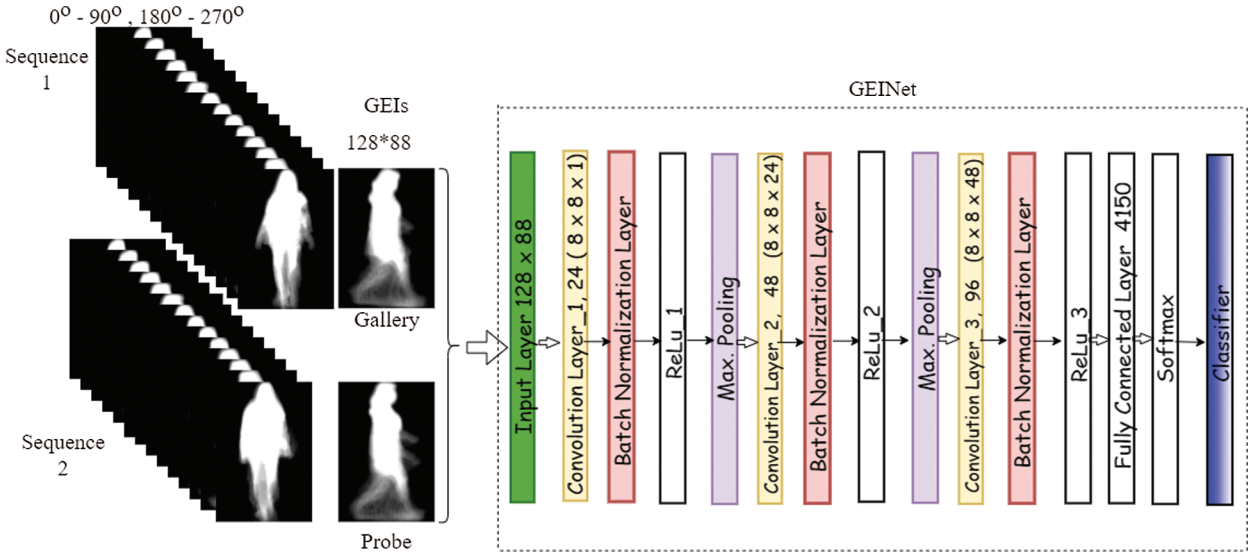

Figure 6: Proposed Improved GEI Network (GEINet)

Table 1: Variable settings for improved GEINet

Variable	Settings
Batch size	64
Momentum	0.9
Maximum Epochs	20
Learning rate schedule	Constant
Number of Iterations	68000
Initial learning rate	0.01
Validation Frequency	30

Table 2: Layer configurations for the improved GEINet

Layer/ Parameter	Parameter Description	No. of Kernels	Size/stride/ padding	Activation function	Pooling
Conv_1	First Convolutional Layer	24	8 × 8 × 1/1/1	ReLu	
Batch Norm.	Batch Normalization Layer		24		
Pooling_1	First Pooling Layer		2 × 2/2/0		Max. pooling
Conv_2	Second Convolutional Layer	48	8 × 8 × 1/1/1	ReLu	
Batch Norm.	Batch Normalization Layer		48		
Pooling_2	Second Pooling Layer		2 × 2/2/0		Max. pooling
Conv_3	Third Convolutional Layer	96	8 × 8 × 1/1/1	ReLu	
Batch Norm.	Batch Normalization Layer		96		

Batch normalization as it differs from other neural network blocks is that it performs computation across images/feature maps in a batch. y = vl_nnbnorm (x, w, b) normalizes each channel of the feature map x averaging over spatial locations and batch instances [30]. Let T be the batch size; then

$$x,\ y\ \in\ R^{H\times W\times K\times T},\ w\ \in\ R^{k},\ b\ \in\ R^{k} \tag{1}$$

The ReLU operator can be expressed in matrix notation as [30]

$$\text{vec } y\ =\ \text{diag s vec x},\ \frac{dz}{d\ vec\ x}\ =\ \text{diag s}\ \frac{dz}{dvec\ y} \tag{2}$$

where $s\ =\ [\text{vec x}\ >\ 0] \in \{0,\ 1\}^{HWD}$ is an indicator vector.

2.4 Setup for Improved GEINet

The sample data were loaded into the image datastore. Image datastore labels the images automatically according to folder names and stores the data as an Image Datastore object. An image datastore enabled us to stock huge image data, comprising data that could not be suitable in memory, and efficiently read batches of images during the training of a convolutional neural network. We trained using the stochastic gradient descent with momentum (SGDM) optimizer with a 0.01 learning rate. Fig. 4 shows the training progress plot, mini-batch loss, and accuracy as well as the validation loss and accuracy. The pre-trained deep convolutional networks available are mostly trained using RGB images which made the networks have three channel inputs automatically. It is therefore not suitable for transfer learning using grayscale single-channel GEI images [29]. Proposed a network call GEINet that is suitable for the GEI image. To improve this network, we added a batch normalization layer to normalize the activations and gradients propagating over a network. The network configuration is available in Tab. 2. A total of 28 samples from two

sequences of GEIs are used. 24 were selected at random for training while the remaining 4 for validation and testing. Only 4150 subjects were found to have balance training samples that are, having both gallery and probe samples in the entire dataset after unzipping. Similarly, [25] pointed out the missing of hundred GEIs which resulted in an unbalanced training sample from the OULP dataset.

A subcategory of the OU-ISIR multi-view large population dataset was used for training and testing [27]. The subset is composed of two gait image sequences from 4150 subjects, 4150 from gallery sequence for training, and 4150 from probe sequence for testing. Fig. 5 shows the GEI images from the OU-ISIR multi-view large population dataset with 0–90 and 180–270° observation angles. The number of frames per subject within the gait cycle ranges from 18 to 35 frames.

The total number of iterations was 5160, with 680 iterations per epochs. The improved GEINet was trained and tested using Matlab 2017b on an NVIDIA Aorus GeForce GTX 1080Ti 11GB GDDR5X 352bit.

The sample data were loaded into Image Datastore. Image Datastore labels the images automatically according to folder names and stores the data as an Image Datastore object. An image datastore enabled us to stock huge image data, comprising data that could not be suitable in memory, and efficiently read batches of images during the training of a convolutional neural network. We trained using the stochastic gradient descent with momentum (SGDM) optimizer with a 0.01 learning rate.

2.5 Preprocessing and Silhouette Frames Normalization

As the frames obtained are 1280 × 960 size as shown in Fig. 7, it is, therefore, necessary to process and normalize it before feeding it to a novel CAMNet. An algorithm in Fig. 8 was developed using the Region of Interest RoI pooling technique to find out all the possible places where the silhouette is located and return the list of bounding boxes corresponding to the position. A tolerance of not more than twenty pixels was added to the left and right before extracting the required object. As millions of frames were involved, it is necessary to automate the process. Image batch processor application in Matlab was employed to execute the function containing the normalization algorithm. Fig. 7 also shows the processed and normalized frames of 128 × 64.

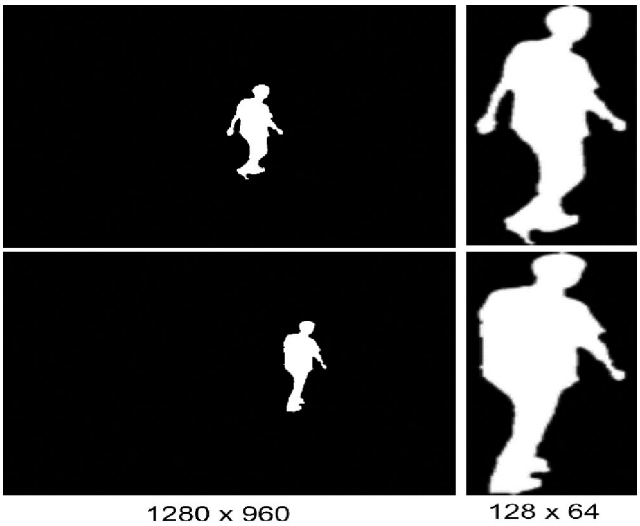

Figure 7: Normalized silhouette at angle 225°

```
Input Read Frame
Define Frame > 100
Convert GrayScale Pixels to Binary Image
Get all the 8-connected pixels found
Measure BoundingBox and Area
 FOR each Length of Measure do
   Get left column, top line, width, height of Frame
   Frame Cropping
   Frame Resizing
 ENDFOR
```

Figure 8: Pseudo-code of frame preprocessing algorithm

2.6 Silhouette Framed Organization and Gait Period Detection

The gait period information for GEI as obtained in section 2.1.2.3 was released to us in addition to the dataset. This information was used as a guide in developing the algorithm that extracts the gait period of a subject within a full walking sequence in the dataset. The program was written in C# to fetch the required gait period and organized the dataset. Fig. 9 shows a normalized gait period of a subject walking at angle 225° containing 26 frames. The number of frames differs between one subject and another.

Figure 9: An illustration of normalized gait period at angle 225°

2.7 DeepGait Feature Extraction

The processes for the DeepGait feature extraction are explained in the following subheadings. Meanwhile, the network activations for a specific layer with additional options specified by one or more name-value pair arguments are returned in a form of features as:

$$h \text{ x } w \text{ x } c \text{ x } n \tag{3}$$

array, where h, w, and c are the height, width, and number of channels for the output of the chosen layer, n is the number of observations.

2.7.1 Maximum Activation Layer Localization and Channel of Interest Pooling

For the localization to be done, we kept track of the maximum activated channel within the feature maps at the third ReLu (ReLu_3) layer in the novel CAMNet shown in Fig. 17. A bounding box was used to highlight the maximum activated channel. For visualization purposes, the maximum activated channels of convolution layer one (Conv_1), convolution layer three (Conv_3), and ReLu layer three are shown in Fig. 10, 12 and 14 respectively in a green bounding box. Fig. 11 shows an input silhouette frame and the extracted maximum activated channel output of convolution layer one (Conv_1). Similarly, Figs. 13 and 15 show the extracted maximum activated channel outputs of convolution layer three (Conv_3) and ReLu layer three (ReLu_3) respectively. It can be seen that, the deeper layer extracts high-level features. The ReLu layer removes all the negative pixels from the frames within a gait period as they are being forwardly propagated through the network.

Figure 10: An illustration of feature maps at convolution layer one with maximum activated channel

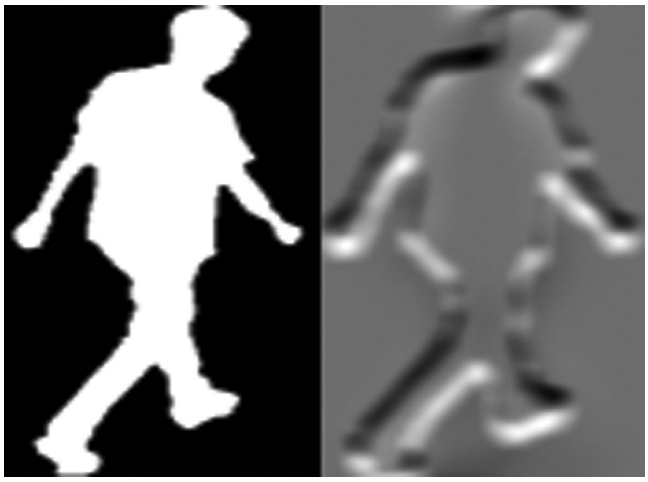

Figure 11: An illustration of input silhouette frame and maximum activated channel output convolution layer one

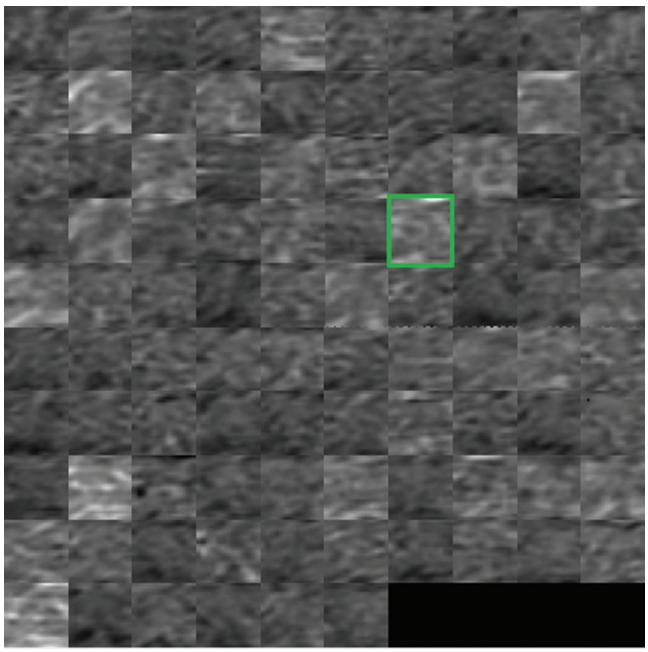

Figure 12: An illustration of feature maps at convolution layer 3 with maximum activated channel highlighted

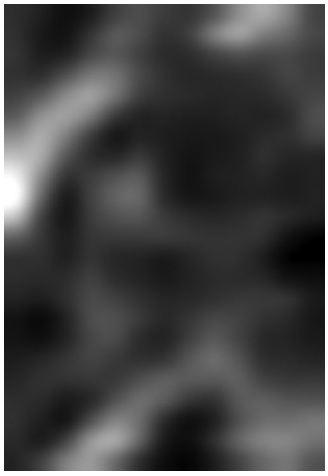

Figure 13: An illustration of maximum activated output frame at convolutional layer 3

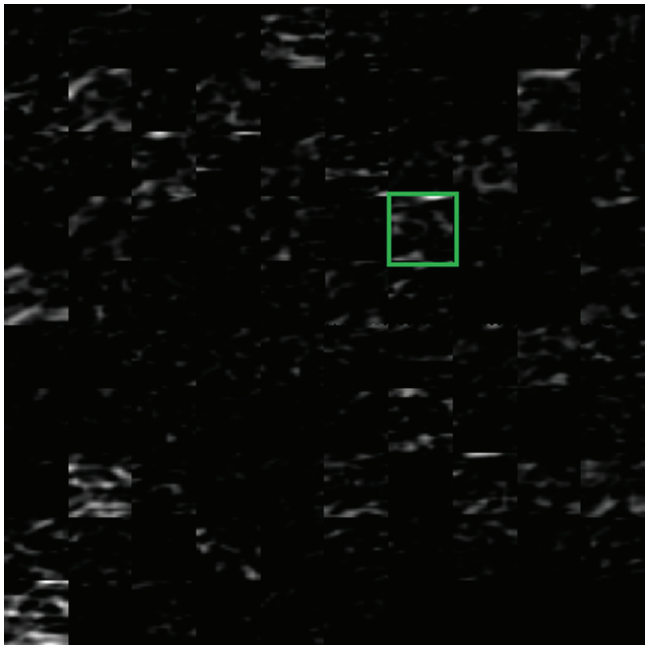

Figure 14: An illustration of feature maps at ReLu layer 3 with maximum activated channel highlighted

The fully connected layer's input size in Fig. 6 is specified as a positive integer or 'auto'. If the input size is 'auto', then the software automatically determines the input size during training or feature extraction. The output size can have n number (4150 in the case of improved GEINet in Fig. 6) of outputs after multiplying the input by a weight matrix and then adds a bias vector. The CAMNet extracts only the maximum activated channel from the feature maps as shown in Fig. 15 and drops the remaining low activated channels (unwanted channels). In essence, the network acts as a maximum channel pass filter after the ReLu_3 layer. The maximum activated channel extracted is then fed to the input of the Spatio-temporal DeepGait feature extraction stage as shown in Fig. 17.

Figure 15: An illustration of a maximum activated output frame of ReLu layer 3

2.7.2 Spatio-temporal DeepGait Feature Extraction

As stimulated by gait energy image (GEI) is obtained by simply taking the average of the silhouette sequence over one gait period which can obtain both the spatial and temporal information [26]. We applied a max-pooling scheme over one gait period of the maximum activated channel to chain the Spatio-temporal information of each subject. Fig. 17 shows the 375-dimensional DeepGait obtained using the OU-MVLP dataset without applying dimension reduction.

Moreover, once the chain of the Spatio-temporal features (DeepGait representation) of each subject is extracted, it will then be used to train the classifier as shown in Fig. 17.

3 Experimental Results

3.1 Overview

We validated the efficiency of the proposed CAMNet during cross-view gait recognition tasks using the OU-MVLP dataset. Two settings are considered: uncooperative subjects' gallery and cooperative subjects' gallery. Views varied among each subject in the uncooperative setting while it is the same for a cooperative setting. The performances of the CAMNet will be compared with the state-of-the-art techniques.

3.2 Representation Generalization and Visualization

As stimulated by gait energy image (GEI) is obtained by simply taking the average of the silhouette sequence over one gait period which can obtain both the spatial and temporal information [26]. The max-pooling scheme was employed over one gait period of ReLu_3 features to chain the Spatio-temporal information of a subject. Another version of ReLu_3 features with average-pooling was tested in our experiments and exhibited poorer performance, which suggests the DeepGait obtained using the CAMNet is valid. Fig. 16 shows the 375-dimensional DeepGait obtained using OUMVLP dataset without applying dimension reduction.

In the a-th gait period, if there are T silhouette images, we can generate T ReLu_3 features. The bth deep convolutional gait representation (DeepGait) element of 375-dimensional representation can then be created by maxing the ReLu_3 features by using Eq. (4).

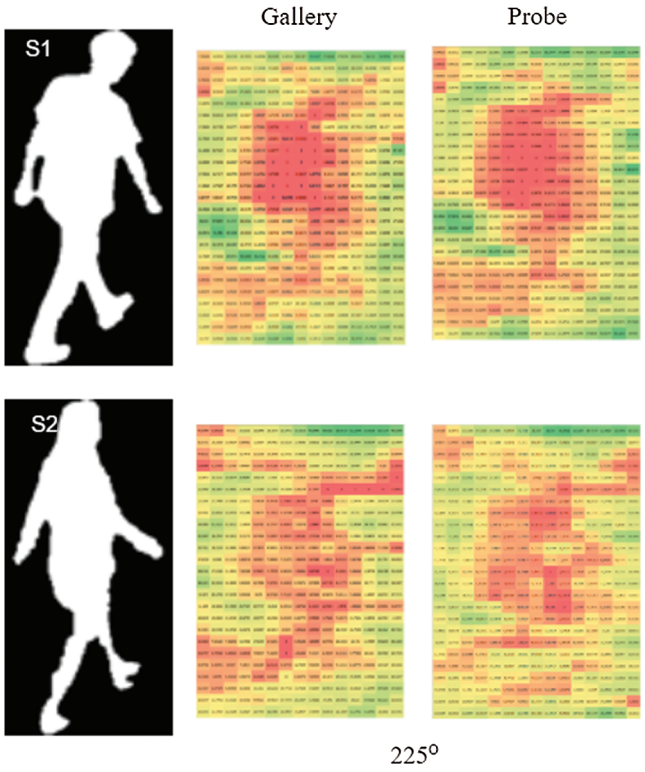

Figure 16: Example of the 375-dimensional DeepGait with 225-degree view angle. S1 and S2 represent two different subjects, separately. We rearrange the vector as 25 × 15 matrices for the convenience of visualization. Approximately 10% of features are non-zero values. Different colors stand for dissimilar values

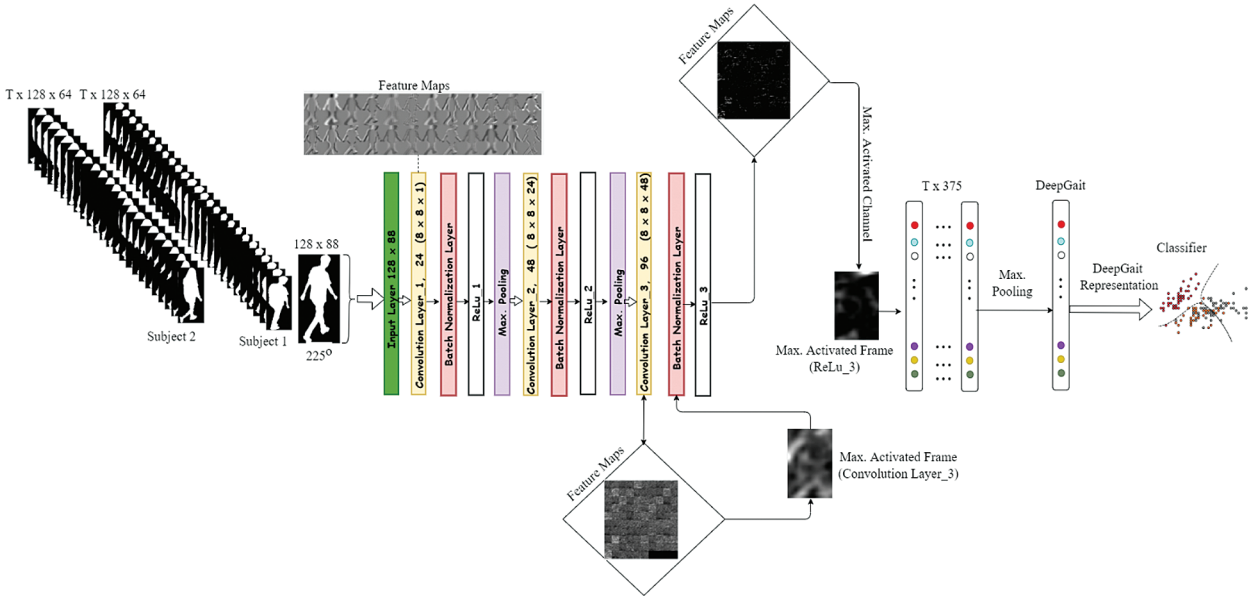

Figure 17: Proposed Novel Channel Activation Map Network (CAMNet)

$$\text{DeepGait}_{a,\,b} = \max_{k=0}^{T-1} \text{ReLu\_3}_{a,\,b,\,k} \tag{4}$$

3.3 Gait Recognition

Gait recognition involves two major tasks: gait verification and gait identification [26]. Gait verification verifies whether two input gait sequences (Gallery, Probe) are for the same subject. In this paper, we calculated a similar score using crossentropyex for the CNN experiment to evaluate the similarity of two given sequences. Whereas, Euclidean distance and spearman were used to evaluate the similarity of two given sequences for the KNN experiment. In gait identification, subjects are gathered (The gallery), and the aim is to decide which of the gallery identities are similar to the probe, the improved GEINet was trained using the gait identification settings. The improved GEINet is shown in Fig. 6.

3.3.1 Results for a Cooperative Setting

We evaluated the recognition accuracy of CAMNet KNN, for all possible view angles (14 probe views vs. 14 gallery views) as shown in Tab. 3. Taking a close look at the approaches, we noticed that CAMNet KNN yielded the best accuracy for the verification task. KNN is selected because of the fact it yielded better accuracy than the other classification algorithms such as SVM. Ten distance metrics of the KNN classifier were used to evaluate the similarity of two given sequences using the DeepGait features and the recognition accuracies are shown in Fig. 18.

Table 3: Recognition accuracy of CAMNet-KNN for all view angle pairs for a cooperative setting

Angles (degree)	1in-GEINet[1]	Improved GEINet[2]	CAMNet KNN[3] (EUCLIDEAN DISTANCE)	CAMNet KNN[4] (SPEARMAN)
0-0	75.9	68.3	72.0(±0.8)	75.9(±0.8)
15–15	87.4	79.6	81.7(±0.8)	85.8(±0.8)
30–30	89.6	84.1	81.4(±0.8)	84.2(±0.8)
45–45	89.3	88.2	81.5(±0.8)	84.7(±0.8)
60–60	86.5	83.4	65.5(±0.8)	71.1(±0.8)
75–75	88.8	85.5	77.5(±0.8)	81.0(±0.8)
90–90	90.7	86.9	81.2(±0.8)	87.1(±0.8)
180–180	83.9	80.8	82.5(±0.8)	**90.1**(±0.8)
195–195	89.9	84.9	87.7(±0.8)	**94.0**(±0.8)
210–210	91.8	86.3	89.4(±0.8)	**94.1**(±0.8)
225–225	91.2	87.4	89.0(±0.8)	**93.4**(±0.8)
240–240	88.5	84.4	73.0(±0.8)	79.1(±0.8)
255–255	89.3	83.4	76.6(±0.8)	75.5(±0.8)
270–270	89.6	84.7	79.2(±0.8)	81.3(±0.8)

1 [27]
2 Proposed Improved GEINet using CNN
3 Novel CAMNet using Euclidean distance
4 Novel CAMNet using Spearman

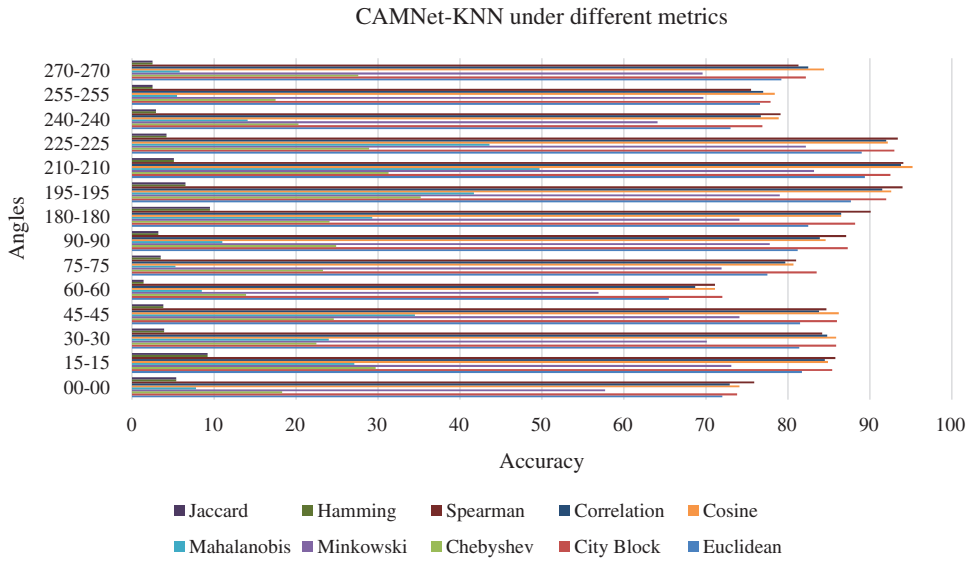

Figure 18: The comparison of the recognition accuracy of CAMNet-KNN under different metrics

Moreover, to specify the angles that have a significant impact on our experiments, we will analyze the percentage of accuracies of each of these angles. It can be seen from Tab. 3 which, the angles with significant impact are straight-line angles and also angles within the reflex angle. The straight-line angle has an accuracy of 90.1% with a tolerance of 0.8 on the spearman distance metric while the angles within the reflex angle which are 195°, 210°, and 225° have the accuracies of 94%, 94.1%, and 93.4% with a tolerance of 0.8 on a spearman distance metric. In essence, these significant impacts might be because adequate features of the subject are revealed from the back compared with inadequate features captured at acute related angles which are the front. To protect against overfitting, 50% hold out was used on the dataset during the training.

3.3.2 Results for an Uncooperative Setting

We evaluated the recognition accuracy of CAMNet KNN for uncooperative setting on the 14 view angles (0°, 15°, ···, 270°) for each subject. According to Fig. 19 and Tab. 4, the results have almost the same tendency as the cooperative setting. Comparing the accuracy of the cooperative setting with that of the uncooperative setting using all the 14 view angles shows that the uncooperative setting has the worse accuracies. However, one of the reasons for this performance degradation is the uncooperative gallery. The slight performance degradation of the uncooperative setting is described as the difference between probe and gallery DeepGaits of the same subject. The difference in view angles is sometimes larger than that of different subjects with the same view angle, which is an indispensable challenge of the uncooperative setting.

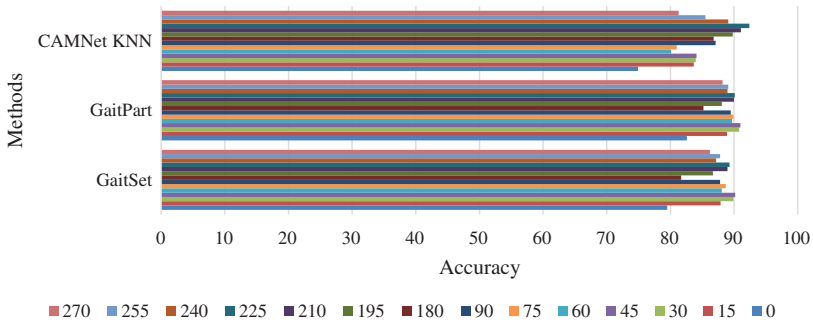

Figure 19: The comparison of the recognition accuracies of CAMNet-KNN with GaitPart and GaitSet

Table 4: Recognition accuracy for uncooperative setting using all the 14 view angles excluding probe angle

Probe	Gallery of 14 view angles		
	GaitSet[5]	GaitPart[6]	CAMNet KNN[7]
0°	79.5	82.6	74.9(±0.8)
15°	87.9	88.9	83.7(±0.8)
30°	89.9	90.8	84.0(±0.8)
45°	90.2	91.0	84.1(±0.8)
60°	88.1	89.7	80.1(±0.8)
75°	88.7	89.9	81.0(±0.8)
90°	87.8	89.5	87.1(±0.8)
180°	81.7	85.2	**86.8**(±0.8)
195°	86.7	88.1	**89.8**(±0.8)
210°	89.0	90.0	**91.1**(±0.8)
225°	89.3	90.1	**92.4**(±0.8)
240°	87.2	89.0	89.1(±0.8)
255°	87.8	89.1	85.5(±0.8)
270°	86.2	88.2	81.3(±0.8)

5 [10]
6 [11]
7 CAMNet using KNN

4 Discussion

We evaluated the recognition accuracy for the verification task that is, one-to-one matching using the proposed improved GEINet, and obtained competitive accuracies when compared with 1in-GEI [27] trained on over 10,000 subjects as shown in Tab. 3. Also, we validated the proposed CAMNet on the state-of-the-art gait dataset for the verification task. The results in Tabs. 3 and 4 show that rear angles of 180, 195, 210, and 225 have higher accuracy than other view angles, which might be because adequate features of the subject are revealed from the back compared with inadequate features captured at acute related angles which are the front.

It can be seen from Tab. 3 that CAMNet KNN (spearman) outperformed the 1in-GEINet CNN on these rear angles of 180, 195, 210, and 225 for the cooperative setting. Moreover, the CAMNet KNN (Euclidean distance) competes with 1in-GEINet CNN [27] on the same rear angles. This can be because more features are available from the back view with more joint information than the front view. DeepGait generated and the processed silhouette would be made available for easy replicability and further research subject to the approval of the dataset providers.

Furthermore, CAMNet KNN (spearman) was also compared with [10] and [11] and outperformed them with rear angles of 180, 195, 210, and 225 for the uncooperative setting. It also competes favorably on the other view angles as shown in Tab. 4.

5 Conclusion

This paper defined a scheme of gait recognition using a novel CAMNet based on CNN. To be explicit, we designed CAMNet that comprises three sequential triplets of convolution, pooling, and batch normalization

layers, and an external maximum pooling to capture the Spatio-temporal information of multiple frames in one gait period. Euclidean distance and spearman metrics based on KNN that outputs a set of similarities to individual training subjects given a DeepGait as input were used as they exhibited better accuracies than other classification algorithms such as SVM. As a result of experiments for cross-view gait recognition in both cooperative and uncooperative settings using the OU-ISIR multi-view large population dataset, we confirmed that, the proposed scheme significantly outperformed the state-of-the-art approaches using this dataset at the rear angles of 180, 195, 210, and 225. A path for future research is to validate the proposed novel method with a different dataset and on a different covariate like the carrying condition.

Acknowledgement: In this paper, we used the OU-ISIR gait database, a multi-view large population dataset (OU-MVLP). We, therefore, thank them for their contribution.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[2] T. H. W. Lam, K. H. Cheung and J. N. K. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recognition*, vol. 44, no. 4, pp. 973–987, 2011.

[3] K. Bashir, T. Xiang and S. Gong, "Gait recognition without subject cooperation," *Pattern Recognition Letters*, vol. 31, no. 13, pp. 2052–2060, 2010.

[4] D. Muramatsu, Y. Makihara and Y. Yagi, "View transformation model incorporating quality measures for cross-view gait recognition," *IEEE Trans. on Cybernetics*, vol. 46, no. 7, pp. 1602–1615, 2016.

[5] S. Yu, H. Chen, E. B. García and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, Hawaii, USA, 87, pp. 532–539, 2017.

[6] S. Yu, R. Liao, W. An, H. Chen, E. B. Garcia *et al.,* "GaitGANv2: Invariant gait feature extraction using generative adversarial networks," *Pattern Recognition*, vol. 87, no. 1, pp. 179–189, 2019.

[7] S. Yu, H. Chen, Q. Wang, L. Shen and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, no. 2, pp. 81–93, 2017.

[8] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu *et al.,* "Gait recognition via disentangled representation learning," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 4705–4714, 2019.

[9] X. Li, Y. Makihara, C. Xu, Y. Yagi and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and covariate features," in *Proc. 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 13306–13316, 2020.

[10] H. Chao, Y. He, J. Zhang and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," *Proc. 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conf., IAAI, 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, vol. 33, no. 16, pp. 8126–8133, 2019.

[11] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou *et al.,* "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE Computer Society Conf. on Computer Vision Pattern Recognition*, Seattle, WA, USA, pp. 14213–14221, 2020.

[12] C. Xu, Y. Makihara, X. Li, Y. Yagi and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, no. 8, pp. 1–15, 2020.

[13] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3953, pp. 151–163, 2006.

[14] S. Zheng, J. Zhang, K. Huang, R. He and T. Tan, "Robust view transformation model for gait recognition," in *Proc. ICIP*, Brussels, pp. 2073–2076, 2011.

[15] W. Kusakunniran, Q. Wu, H. Li and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," in *Proc. of IEEE 12th Int. Conf. on Computer Vision Workshops*, ICCV Workshops, Kyoto, pp. 1058–1064, 2009.

[16] X. Xing, K. Wang, T. Yan and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recognition*, vol. 50, no. 19, pp. 107–117, 2016.

[17] Z. Zhang, J. Chen, Q. Wu and L. Shao, "GII Representation-based cross-view gait recognition by discriminative projection with list-wise constraints," *IEEE Trans. on Cybernetics*, vol. 48, no. 10, pp. 2935–2947, 2018.

[18] F. M. C. Payanores, "Gait recognition from multiple view-points," Ph.D. dissertation. Department of Computer Architecture, University of Córdoba, Córdoba, Spain, 2018.

[19] P. P. Min, S. Sayeed and T. S. Ong, "Gait recognition using deep convolutional features," in *Proc. of 7th Int. Conf. on Information and Communication Technology (ICoICT)*, Kuala Lumpur, Malaysia, pp. 1–5, 2019.

[20] Q. Zou, Y. Wang, Q. Wang, Y. Zhao and Q. Li, "Deep learning based gait recognition using smartphones in the wild," *IEEE Trans. on Information Forensics and Security*, vol. 15, pp. 3197–3212, 2018.

[21] M. Alotaibi and A. Mahmood, "Improved gait recognition based on specialized deep convolutional neural network," *Computer Vision and Image Understanding*, vol. 164, no. 13, pp. 103–110, 2017.

[22] F. Battistone and A. Petrosino, "TGLSTM: A time based graph deep learning approach to gait recognition," *Pattern Recognition Letters*, vol. 126, no. 6, pp. 132–138, 2019.

[23] F. Horst, S. Lapuschkin, W. Samek, K. R. Müller and W. I. Schöllhorn, "Explaining the unique nature of individual gait patterns with deep learning," *Scientific Reports*, vol. 9, no. 1, pp. 255, 2019.

[24] S. Tong, Y. Fu, X. Yue and H. Ling, "Multi-view gait recognition based on a spatial-temporal deep neural network," *IEEE Access*, vol. 6, pp. 57583–57596, 2018.

[25] Z. Wu, Y. Huang, L. Wang, X. Wang and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 209–226, 2017.

[26] C. Li, X. Min, S. Sun, W. Lin and Z. Tang, "DeepGait: A learning deep convolutional representation for view-invariant gait recognition using joint Bayesian," *Applied Sciences*, vol. 7, no. 3, pp. 1–15, 2017.

[27] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. on Computer Vision and Applications*, vol. 10, no. 1, pp. 882, 2018.

[28] H. Iwama, M. Okumura, Y. Makihara and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 5, pp. 1511–1521, 2012.

[29] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. 2016 Int. Conf. on Biometrics (ICB), Halmstad*, pp. 1–8, 2016.

[30] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for matlab," in *Proc. of 23rd ACM Int. Conf. on Multimedia*, New York, USA, pp. 689–692, 2015.