IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 12, 2021, accepted July 29, 2021, date of publication August 6, 2021, date of current version August 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3102936

Gait Recognition and Re-Identification Based on Regional LSTM for 2-Second Walks

PIYA LIMCHAROEN, NIRATTAYA KHAMSEMANAN, AND CHOLWICH NATTEE
Sirindhorn International Institute of Technology, Thammasat University, Khlong Luang, Pathum Thani 12120, Thailand

Corresponding authors: Nirattaya Khamsemanan (nirattaya@siit.tu.ac.th) and Cholwich Nattee (cholwich@siit.tu.ac.th)

This work was supported by the Thammasat University Research Unit in Gait Analysis and Intelligent Technology (GaitTech).

ABSTRACT Law enforcement and different authorities need a new efficient way to track and re-identify a person of interest via different cameras. Usually, the person of interest is not known and the original video may be short and have poor quality. In this paper, we propose a new technique based on a new regional-LSTM learning model that can use a 2-second walk to recognize and re-identify an unknown person. The proposed technique first targets the rhythm of movements in different regions of the body by creating a separate LSTM model for each region. Then, outputs from 22 regions are combined in a subnetwork to extract the relations and different degrees of uniqueness of all regions. The proposed regional LSTM model creates a gait-embedded vector to represent a 2-second walk. Experimenting on imbalanced and balanced datasets, the results show that the proposed regional LSTM model performs significantly better than the existing techniques on the Cumulative Matching Characteristic (CMC) curves and top-k accuracy, Receiver Operating Characteristic (ROC) curves, and Precision-Recall (PR) curves. This indicates that the proposed technique has a high-ranking performance (CMC test), can efficiently distinguish the gaits of a subject from others (ROC test), and occupies high relevancy (PR test). From the experimental results, it is likely that one in four videos retrieved from the proposed techniques shows the person of interest with over 90.8% and 85.7% accuracies in imbalanced and balanced data, respectively. This demonstrates that the proposed regional LSTM model is efficient and useful in tracking and re-identifying a person of interest.

INDEX TERMS Gait biometric, gait recognition, gait verification, gait-embedded vector, human recognition, information retrieval, re-identification, regional LSTM, separate LSTM, similarity ranking, softmax loss.

I. INTRODUCTION

Gait recognition and re-identification have an important role in the fields of security, authentication, surveillance, and commercials. They can be utilized from afar without subjects' awareness [1]–[3]. This sets gait recognition apart from traditional biometric recognition techniques such as signature, retinal, or facial recognition.

Gait is a behavior biometric property of a person. It is a locomotion pattern of a living organism. In layman's term, gait is the way a person walks and move around in his or her natural behavior. Most people have their unique ways of walking. As a consequence, gait and physical biometric properties, such as limb lengths, can be used to identify a person with high accuracies.

Gait recognition is a technique to identify a person from the way they walk. A re-identification is a process of associating images or videos of the same person from different angles, cameras, and occasions.

Conventional gait recognition techniques are supervised. This means that the gaits of known subjects are required in order for a gait recognition to identify a person. Conventional gait recognition techniques are divided into two categories: model-free and model-based [1], [4], [5].

In model-free (or appearance-based) gait recognition techniques, gait characteristics or gait features are extracted by separating a person's appearances from the background. These appearances are converted into gait features, such as contours, silhouettes, and depth, in the model-free techniques. In early works of model-free approaches, gait recognition was focused on features such as silhouettes and contours. However, the accuracies of model-free gait

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

recognition techniques depend on silhouette qualities. The qualities of silhouettes are associated with a subject's environment or a subject's conditions, such as a subject's movements, directions, clothing, camera viewpoint, walking surface, and lighting environment, as discussed in [1], [4], [6], [7]. A number of model-free approaches [6]–[11] have been proposed to handle the viewpoint issue. In [12], a GEI (Gait Energy Image) is introduced as a gait feature. A GEI is an image that is created by combining all spatial-temporal silhouettes of a subject's walking within a limited duration into a 2D image. The GEI is widely used in many model-free approaches [6], [7], [9], [11], [13]. Some model-free gait recognition works [14], [15] used the LSTM on model-free datasets (sequential silhouette images).

In model-based approaches, gait data is simplified into a known structure before the feature extraction processes. These structures mimic human skeletons or human body structures. Recently, many gait recognition works have focused on model-based approaches because of new technologies, such as a Microsoft Kinect, which make it easier to construct human skeletons from videos. A Microsoft Kinect was initially designed as an input interface for a gaming device (XBOX-360). A Kinect and its SDK generate a 3D skeleton stream output directly from a video stream. A skeleton stream is a series of frames where a subject's body joints are represented as points in 3D space in a frame. Many model-based gait recognition techniques are based on skeleton data obtained from Kinect devices.

Some model-based gait recognition works, [16]–[18], use mainly body structures (static features) obtained from Kinect devices, such as limb lengths and body heights as features. However, these techniques do not achieve high accuracy with datasets that contain subjects with similar body builds. In contrast, some model-based gait recognition techniques, [19]–[29], use body movements (dynamic features) such as stride lengths or arm movements with body structures as gait features. These techniques perform better than gait recognition techniques that mainly use static features. Some existing model-based works [5], [30]–[32] have operated the recognition process based on the MLP or the CNN on the walk's slices (a frame or a few frames) without using the benefits of sequential data. However, most of these techniques do not withstand situations where gait data are collected from different observational viewpoints. This issue is, sometimes, referred to as a view-point issue.

Andersson *et al.* [19] propose a gait recognition technique that uses a combination of mostly static with a few dynamic gait features for a fixed-direction walk (view-dependent). The static features used in their work are limb lengths and heights. Dynamic features used in their work are standard deviations and means of angles between limbs, stride lengths, and velocities. Andersson *et al.* [19] conducted experiments on a dataset collected from 140 subjects where the subjects were asked to walk in a semi-circular direction while the camera continuously captured a walk from sideways (fixed-direction walk). Their work shows the highest accuracy of 87.7% based on using the k-NN algorithm with Manhattan distance as the distance function and a parameter ($k = 5$) in the classification process.

In [20], Yang *et al.* propose another gait recognition technique using some static and more dynamic features than those used in [19]. A walk is represented in the form of a vector, similar to the technique presented in [19]. In the classification process, the technique from [20] uses k-NN with the Manhattan distance, similar to [19]. On the same dataset collected by [19], Yang's technique yields an accuracy of 95.4% which outperforms [19]. The results confirm that static features alone are not enough to achieve high accuracies. Dynamic features are needed to accomplish that goal. However, both [20] and [19] do not provide a way to handle different viewpoint issues.

Many gait recognition techniques, [17], [21], [23]–[25], [27], design their algorithms based on gait data collected from an entire walk, which usually lasts 15 seconds or more (walk-based). In these techniques, a gait feature such as the stride length of an entire walk, are represented as a few numbers, such as a mean and a standard deviation, which may be too simple. Some unique characteristics of gait get lost in the process.

Recently, cycle-based [28], [29] and frame-based [5], [30], [33] gait recognition techniques have been proposed. In these techniques, a walk is split into smaller slices (either walking cycles or a certain number of consecutive frames). Gait features are extracted from each smaller slice. Gait features of a walk are represented in more complicated forms than just a few numbers or a single vector. Cycle-based and frame-based gait recognition techniques outperform walk-based gait recognition techniques. This suggests that gait features, extracted from cycle-based and frame-based gait recognition techniques, contain more unique information about a person.

In facial recognition fields, many techniques [34]–[44] use deep neural network learning (metric learning) to construct a facial embedding. The metric learning is a problem of learning to transform or finding a representation function that maps the input feature into an embedded space (or create an embedding output), whose the variation of intra-class identity is small, but the variation of inter-class identity is large. A facial embedding is a vector that represents the features extracted from the face. Many of these networks are generalized enough. The models can be used to construct a face embedding without requiring a label (unsupervised). Different loss functions such as Softmax loss, contrastive loss, triplet loss, center loss, feature and weight normalization, and large margin loss, are used to enhance the discriminative power between the clusters of face embeddings [45].

In this work, we propose a new gait recognition and re-identification technique that is unsupervised, resists the view-point issue, and only requires a 2-second walk input. The overview process of the proposed technique is shown in Fig. 1. We propose a new unsupervised gait recognition technique based on a new learning model, called the regional-LSTM learning model. The regional-LSTM

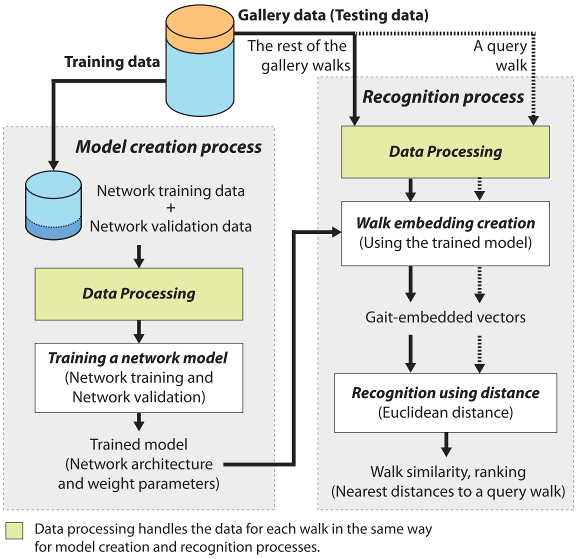
Gallery data (Testing data)
Training data
The rest of the gallery walks
A query walk
Recognition process
Model creation process
Network training data + Network validation data
Data Processing
Training a network model (Network training and Network validation)
Trained model (Network architecture and weight parameters)
Data Processing
Walk embedding creation (Using the trained model)
Gait-embedded vectors
Recognition using distance (Euclidean distance)
Walk similarity, ranking (Nearest distances to a query walk)
Data processing handles the data for each walk in the same way for model creation and recognition processes.

FIGURE 1. **Overview process of the proposed technique.**

learning model is a representation function that maps gait features into an embedded space so that the similarity of intra-class is small and the similarity of inter-class is large. The proposed technique focuses on sequential movements of each region of the body by creating an LSTM model for each region. It then combines the outputs from all regions to create a gait-embedded vector for an entire body. By doing this different regions are assigned different weights to reflect different degrees of uniqueness in the regions. An output of the regional-LSTM learning model is an embedded vector in Euclidean space.

The experimental results show that our proposed technique outperforms existing techniques in ranking performance (Cumulative Matching Characteristic (CMC) curves), separability (Receiver Operating Characteristic (ROC) curves), and relevancy (Precision-Recall (PR) curves). The results suggest that the proposed technique is suitable for real-world uses.

This paper is organized as follows. Section II illustrates the construction and architecture of the regional LSTM learning model and the overall process of the proposed techniques. Section III explains the experimental setups and three performance evaluations: CMC curves, ROC curves, and PR curves. The results and discussion of all three performance evaluations are in Section IV. Section V contains the conclusion of this work.

II. METHOD

In this work, an input of our technique is a 2-second walk captured from a Kinect V1 device. A 2-second walk consists of 40 consecutive frames of skeleton data (3-dimensional coordinates of 20 joints). We define walk w as

$$w = \langle F^1, \ldots, F^{40} \rangle, \tag{1}$$

where F^k is frame k, for $k = 1, \ldots, 40$ of walk w. Frame F^k is defined as

$$F^k = \left[\mathbf{f}_l^k\right] = \begin{bmatrix} \mathbf{f}_1^k \\ \vdots \\ \mathbf{f}_{20}^k \end{bmatrix}, \tag{2}$$

where $\mathbf{f}_l^k$ is a 3-dimensional coordinate (x, y, z) of joint l, for $l = 1, \ldots, 20$, in frame k. Walk w has dimensions of 40 (frames) × 20 (joints) × 3 (coordinate x, y, z).

A. DATA PROCESSING

1) NOISE REMOVAL PROCESS

To remove noise from an input, we average the 3-dimensional coordinates of two consecutive frames of the same joint. The averaging process helps to reduce the distortions in the frame's coordinate but still maintains the vital information since only two consecutive terms are averaged. This process also shortens the length of an input by half, and consequently, reduces the overall computational time.

In the noise removal process, the joint coordinates of pairs of two consecutive frames $\{F^1, F^2\}$, $\{F^3, F^4\}, \ldots,$ $\{F^k, F^{k+1}\}, \ldots, \{F^{39}, F^{40}\}$ are averaged. The output of this process is

$$\bar{w} = \langle G^1, \ldots, G^{20} \rangle, \tag{3}$$

where

$$G^i = \frac{F^{2i-1} + F^{2i}}{2}, \tag{4}$$

for $i = 1, \ldots, 20$.

Output $\bar{w}$ has dimensions of 20 (averaged frames) × 20 (joints) × 3 (coordinate x, y, z).

2) REGIONAL REPRESENTATION

Three connected joints in an average frame are grouped into a region, as defined in Table 1. An example of region 1 is shown in Fig. 3. Each region consists of a left joint, a middle joint, and a right joint. For region r of average frame i, we define $c^i_{r,Left}$ as a vector starting from the middle joint and ending at the left joint, and $c^i_{r,Right}$ as a vector starting from the middle joint and ending at the right joint. This structure is similar to Joint Replacement Coordinates (JRCs) [5].

Output w' of this process is defined as

$$w' = \langle C^1, \ldots, C^{20} \rangle, \tag{5}$$

where

$$C^i = \begin{bmatrix} \mathbf{c}^i_{1,Left} & \mathbf{c}^i_{1,Right} \\ \vdots & \vdots \\ \mathbf{c}^i_{r,Left} & \mathbf{c}^i_{r,right} \\ \vdots & \vdots \\ \mathbf{c}^i_{22,Left} & \mathbf{c}^i_{22,Right} \end{bmatrix}, \tag{6}$$

for $i = 1, \ldots, 20$. Output w' has dimensions of 20 (averaged frames) × 22 (regions) × 2 (vectors) × 3 (coordinate x, y, z).

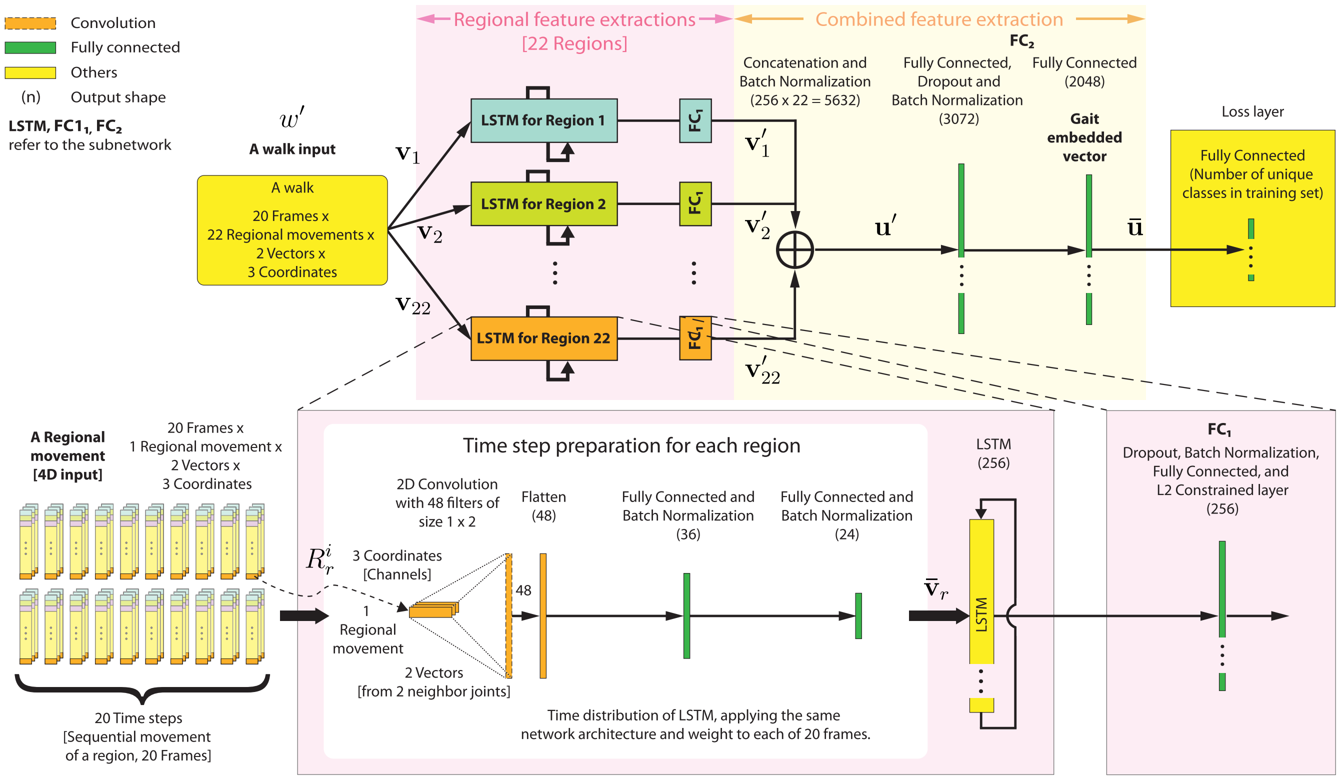
Convolution
Fully connected
Others
(n) Output shape
LSTM, FC1₁, FC₂ refer to the subnetwork
w'
A walk input
A walk
20 Frames x
22 Regional movements x
2 Vectors x
3 Coordinates
$\mathbf{v}_1$
$\mathbf{v}_2$
$\mathbf{v}_{22}$
Regional feature extractions [22 Regions]
LSTM for Region 1
LSTM for Region 2
LSTM for Region 22
FC₁
Combined feature extraction
Concatenation and Batch Normalization (256 x 22 = 5632)
FC₂
Fully Connected, Dropout and Batch Normalization (3072)
Fully Connected (2048)
Gait embedded vector
$\mathbf{v}'_1$
$\mathbf{v}'_2$
$\mathbf{v}'_{22}$
$\mathbf{u}'$
$\bar{\mathbf{u}}$
Loss layer
Fully Connected (Number of unique classes in training set)
A Regional movement [4D input]
20 Frames x
1 Regional movement x
2 Vectors x
3 Coordinates
20 Time steps [Sequential movement of a region, 20 Frames]
R^i_r
Time step preparation for each region
2D Convolution with 48 filters of size 1 x 2
3 Coordinates [Channels]
1 Regional movement
2 Vectors [from 2 neighbor joints]
48
Flatten (48)
Fully Connected and Batch Normalization (36)
Fully Connected and Batch Normalization (24)
$\bar{\mathbf{v}}_r$
LSTM (256)
LSTM
Time distribution of LSTM, applying the same network architecture and weight to each of 20 frames.
FC₁
Dropout, Batch Normalization, Fully Connected, and L2 Constrained layer (256)

FIGURE 2. Proposed network for extracting a regional feature (Regional-LSTM) for joints: Overall network architecture for the Regional-LSTM.

TABLE 1. List of regions.

Region (index r)	Left Joints	Middle Joints	Right Joints
1	Knee Right	Ankle Right	Foot Right
2	Hip Right	Knee Right	Ankle Right
3	Hip Center	Hip Right	Knee Right
4	Spine	Hip Center	Hip Right
5	Knee Left	Ankle Left	Foot Left
6	Hip Left	Knee Left	Ankle Left
7	Hip Center	Hip Left	Knee Left
8	Spine	Hip Center	Hip Left
9	Head	Shoulder Center	Shoulder Right
10	Spine	Shoulder Center	Shoulder Right
11	Shoulder Center	Shoulder Right	Elbow Right
12	Shoulder Right	Elbow Right	Wrist Right
13	Elbow Right	Wrist Right	Hand Right
14	Head	Shoulder Center	Shoulder Left
15	Spine	Shoulder Center	Shoulder Left
16	Shoulder Center	Shoulder Left	Elbow Left
17	Shoulder Left	Elbow Left	Wrist Left
18	Elbow Left	Wrist Left	Hand Left
19	Hip Left	Hip Center	Hip Right
20	Shoulder Left	Shoulder Center	Shoulder Right
21	Shoulder Center	Spine	Hip Center
22	Head	Shoulder Center	Spine

B. REGIONAL-LSTM LEARNING MODEL

We propose a new network architecture design called the Regional-LSTM learning model. The overall structure of the proposed technique is shown in Fig. 2. The proposed model is designed to handle a series (regions) of sequential data, unlike the conventional LSTM model that handles a single sequence.

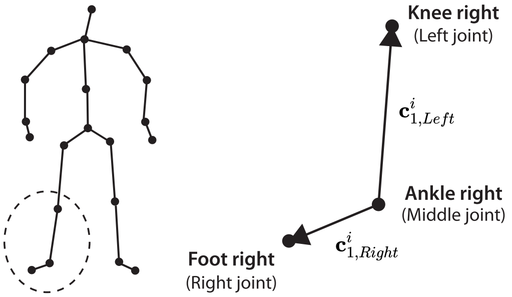

FIGURE 3. Example of joints in region number one (r = 1), showing one region from 22 regions.

1) REGIONAL FEATURE EXTRACTION

For region r, let $\mathbf{v}_r$ be a slice of w' that contains region r. This means that

$$\mathbf{v}_r = \langle R^1_r, \ldots, R^{20}_r \rangle, \tag{7}$$

where

$$R^i_r = \left(\mathbf{c}^i_{r,Left}, \mathbf{c}^i_{r,right} \right) \tag{8}$$

for $i = 1, \ldots, 20$.

Each R^i_r is fed into a convolutional layer with 48 filters of size 1 × 2 and the ReLU activation function. Here, we treat each coordinate of R_r^i as a channel. The output of this layer is flattened and fed into two consecutive fully connected layers with 36 and 24 units (with batch normalization). Each R_r^i is transformed into vector $\bar{R}_r^i$, with 24 values.

A sequence of 20 vectors, $\bar{\mathbf{v}}_r = \langle \bar{R}_r^1, \ldots, \bar{R}_r^{20} \rangle$, is fed into an LSTM with 256 units. The output of this step is then a vector with 256 values.

As shown in Fig. 2, we create 22 subnetworks of the same structure for each region. This part of the network is designed to extract meaningful features from a sequence of each regional movement.

The output from each regional LSTM subnetwork is fed into a fully connected layer, FC_1, with dropout (with parameter 0.3), batch normalization, and L2 constrained normalization. A dropout operation helps to prevent the model from entering an overfitted situation. We adopt an L2 constrained normalization from [35].

For each region r, the output of the FC_1 is vector $\mathbf{v}'_r$ with a dimension of 256.

2) COMBINED FEATURE EXTRACTION

Outputs of all regional feature extraction $\mathbf{v}'_r$, for $r = 1, \ldots 22$ are concatenated into a single vector, $\mathbf{u}$, with a dimension of 5632:

$$\mathbf{u} = \bigoplus_{r=1}^{22} \mathbf{v}'_r, \tag{9}$$

followed by batch normalization. The output $\mathbf{u}'$ of this process is a vector of 5632 dimensions.

The output $\mathbf{u}'$ is then fed into a 2-layer subnetwork, FC_2. The first layer is a fully connected layer with 3072 units, followed by a dropout (parameter 0.3), and batch normalization. The second layer is a fully connected layer with 2048 units. The output $\bar{\mathbf{u}}$ of FC_2 is a vector of 2048 dimensions. This output $\bar{\mathbf{u}}$ is used as a gait embedded vector which is a representation of gait from a 2-second walk.

Note that the activation functions used in the LSTM of all regions, FC_1, FC_2 are a hyperbolic tangent function. This activation function is chosen to avoid vanishing and exploding gradient problems.

In this work, the distance between two embedded gait vectors $\bar{\mathbf{u}}_p$ and $\bar{\mathbf{u}}_q$ is

$$dist\left(\bar{\mathbf{u}}_p, \bar{\mathbf{u}}_q\right) = \|\bar{\mathbf{u}}_p, \bar{\mathbf{u}}_q\|_2, \tag{10}$$

which is an L2 norm in Euclidean space.

To train the proposed gait embedding network, we add a fully connected layer, called the loss layer, into the network. In this layer, the activation function is set to the linear function. We set the target to a one-hot vector denoting a person in the training set. Then, the number of units in this layer is set to the number of unique persons in the training set. The parameters of the network are therefore updated, to make it capable of correctly identifying a person. The output of FC_2 can then be used as a gait-embedded vector to represent a walk.

A number of loss functions have been proposed for this structure of network. We use three loss functions in the experiments:

1) *Softmax Cross-Entropy Loss:* This is the conventional loss function used to train the network for multi-class classification:

$$L_{ce} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{W_{y_i}^\top \bar{\mathbf{u}}_i}}{\sum_{j=1}^{c} e^{W_j^\top \bar{\mathbf{u}}_i}}, \tag{11}$$

where n is the number of walks in the training set, c is the number of unique persons, $\bar{\mathbf{u}}_i$ is the gait-embedded vector when the i-th walk is fed into the network, y_i is the index of the target person, and W_j is the weight matrix of the j-th unit in the loss layer where $j = 1, \ldots, c$.

2) *AMSoftmax Loss [40]:* This is a modification of the softmax cross-entropy loss, to maximize cosine similarity margins between classes. This loss function was originally designed for embedding face images.

$$L_{ams} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{s\left(W_{y_i}^\top \bar{\mathbf{u}}_i - m\right)}}{e^{s\left(W_{y_i}^\top \bar{\mathbf{u}}_i - m\right)} + \sum_{j=1, j\neq y_i}^{c} e^{sW_j^\top \bar{\mathbf{u}}_i}} \tag{12}$$

where s is the user-defined scaling parameter, and m is the user-defined margin.

3) *AdaCos [44]:* This is the cosine-based softmax loss function originally designed for face recognition. This function uses a scaling parameter that is dynamically adapted using the training process.

$$L_{adacos} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{\tilde{s}W_{y_i}^\top \bar{\mathbf{u}}_i}}{\sum_{j=1}^{c} e^{\tilde{s}W_j^\top \bar{\mathbf{u}}_i}} \tag{13}$$

where $\tilde{s}$ is the automatically tuned scaling parameter.

III. EXPERIMENTS

A. DATASETS

We used the combined gait data from three different datasets, SIIT-CN-A (91 users), SIIT-CN-B (393 users), and SIIT-CN-C (130 users) [5], which are collected by CN (Cholwich-Nirattaya) Lab. In all three datasets, participants were asked to walk freely, in any direction where multiple Kinects were placed at different heights and angles. Each captured video is at least 15 seconds long.

We create a new gait dataset, called SIIT-CN-D, from these three datasets by splitting the original capture videos into many 2-second (40 frames) videos. The SIIT-CN-D dataset consists of 180,097 2-second walks from 610 random unique subjects with different heights, weights, and genders. These 2-second walks are captured by different camera angles and heights from the ground.

B. EXPERIMENTAL SET-UP

We experiment using the 10-fold cross-validation technique where data in SIIT-CN-D is divided into 10 groups equally, based on unique subjects. Nine groups are used as a training set and the last group is used as a test set. This setup is constructed in this manner so that no subjects in the test set are part of the model training.

We assess the proposed techniques in two different situations: imbalanced and balanced. In an imbalanced situation, all 2-second walks in the test set are used as the gallery. In a balanced situation, 100 2-second walks of each participant are randomly selected to be used as the gallery.

We conduct the following experiments to assess our proposed gait recognition technique, based on the regional LSTM learning model with various loss functions, against three existing techniques, Han *et al.* (GEI) [12], Andersson *et al.* [19], and Yang *et al.* [20]. Most gait recognition techniques are designed to be used in supervised situations (cycle-based and frame-based gait features). Therefore, many of the existing techniques cannot be used in unsupervised situations and experiments. Only a few gait recognition techniques with walk-based gait features, Han *et al.* (GEI) [12], Andersson *et al.* [19], and Yang *et al.* [20], can be used in unsupervised situations. We use gait features proposed in these three existing techniques as gait representation vectors in our experiments. We also conduct experiments using gait data from the entire body as the input to a single LSTM network.

In each fold, experiments are conducted using the leave-one-out technique. A 2-second walk from the gallery is set to be a query walk. For each query, the rest of the gallery is ranked based on the distances. The query walk with the smallest distance is the top rank. For example, a rank-1 walk is a walk in the gallery that has the smallest distance to the query walk. A rank-2 walk is a walk in the gallery that has the second smallest distance to the query walk, etc.

C. PERFORMANCE EVALUATIONS

We conduct the following three evaluations to assess ranking, separability, and relevancy performances of the proposed techniques.

1) CUMULATIVE MATCHING CHARACTERISTIC (CMC) CURVES AND TOP-k ACCURACY

Cumulative Matching Characteristic (CMC) curves and top-k accuracy are a popular assessment for human identification and re-identification. In real-world situations, authorities are interested in a small group of suspects that contains a real criminal. CMC curves and top-k accuracy evaluate the ranking capabilities of a technique. The CMC top-k accuracy is defined as

$$\text{CMC top-}k\text{ accuracy} = \frac{1}{n}\sum_{i=1}^{n}\Phi(i), \tag{14}$$

where, n is the number of (all) samples in the gallery

$$\Phi(i) = \begin{cases} 1, & \text{if the top } k \text{ ranked gallery samples belong} \\ & \text{to the same subject (class) as the query } i; \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

A CMC curve is a graph that plots rank k on the x-axis against CMC top-k accuracy on the y-axis.

2) RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES

A Receiver Operating Characteristic (ROC) curve is a measurement to evaluate how well a technique is capable of distinguishing between classes. ROC curves display the true positive rate on the y-axis and the false positive rate on the x-axis. ROC curves are constructed as follows.

True positive rate k (TPR_k), and false positive rate k (FPR_k) are defined as

$$TPR_k = \frac{\sum_{i=1}^{n} TP_k(i)}{\sum_{i=1}^{n} TP_k(i) + \sum_{i=1}^{n} FN_k(i)}, \tag{16}$$

$$FPR_k = \frac{\sum_{i=1}^{n} FP_k(i)}{\sum_{i=1}^{n} FP_k(i) + \sum_{i=1}^{n} TN_k(i)}, \tag{17}$$

where,

$TP_k(i)$ is the true positives of the top k ranked samples of query sample i, which is the number of samples within top k that belong to the same class as query sample i.

$FP_k(i)$ is the false positives of the top k ranked samples of query sample i, which is the number of samples within top k that do not belong to the same class as query sample i.

$TN_k(i)$ is the true negatives of the top k ranked samples of query sample i, which is the number of samples not in top k that do not belong to the same class as query sample i.

$FN_k(i)$ is the false negatives of the top k ranked samples of query sample i, which is the number of samples not in top k that belong to the same class as query sample i.

A point in an ROC curve is of the form (FPR_k, TPR_k).

An ideal point in an ROC curve is the top left corner where the true positive rate is 100% and the false positive rate is 0% which is not realistic. This means that a technique with a larger area under the ROC curve has higher separability.

3) PRECISION-RECALL (PR) CURVES

A Precision-Recall (PR) curve is a measurement to evaluate the relevancy of a technique. Precision is the ratio of retrieved relevant samples (same class as a query sample) over all retrieved samples (top k samples). Recall is the fraction of retrieved relevant samples over all relevant samples.

PR curves display precision (true positive rate) on the y-axis and recall on the x-axis. PR curves are constructed as follows.

The precision and recall of top k ranked samples are defined as follows.

$$\text{Precision}_k = \frac{\sum_{i=1}^{n} TP_k(i)}{\sum_{i=1}^{n} TP_k(i) + \sum_{i=1}^{n} FP_k(i)}, \tag{18}$$

$$\text{Recall}_k = \frac{\sum_{i=1}^{n} TP_k(i)}{\sum_{i=1}^{n} TP_k(i) + \sum_{i=1}^{n} FN_k(i)}, \tag{19}$$

where, $TP_k(i)$, $FP_k(i)$, $TN_k(i)$, $FN_k(i)$ are as defined in Equations (16) and (17).

A point in a PR curve is of the form $(\text{Recall}_k, \text{Precision}_k)$.

An ideal point in a PR curve is the top right corner where the precision is 100% and the recall is 0%, which is also not realistic. This means that a technique with a larger area under the PR curve has higher relevant capacity.

IV. RESULTS AND DISCUSSION

Results of the CMC top-k accuracy, $k = 1, \ldots, 5$, area under the ROC curves and the area under the PR curves of the proposed techniques and existing techniques, [12], [19], [20], on the imbalanced and balanced galleries are shown in Table 2 and Table 3, respectively.

A. CUMULATIVE MATCHING CHARACTERISTIC (CMC) CURVES AND TOP-k ACCURACY

In reality, a human recognition and re-identification technique is useful to authorities when it could provide a small list of suspects that contains a real criminal. CMC top-k accuracy and curves are tests that are used to measure the ranking performance of human recognition and re-identification techniques. Techniques that perform better in these tests should be the most useful for authorities.

CMC curves of top the 51 ranks of the proposed techniques and existing techniques, [12], [19], [20], and imbalanced and balanced galleries are shown in Fig. 4 and Fig. 5 respectively.

1) IMBALANCED GALLERY

From the experimental results in Table 2 and Fig. 4, the proposed regional LSTM with L2 and Softmax outperform the rest of the techniques on CMC top-k accuracies from $k = 1, \ldots, 5$ significantly except the proposed technique with Softmax. The proposed regional LSTM with L2 and Softmax provides more than 90% accuracy for rank 4. This means that, in a group of 4 suspects that were provided by this technique, it is likely that a real criminal is one of four with 90% accuracy from a 2-second walk. However, the proposed regional LSTM with Softmax also performs well and is not significantly different from the regional LSTM with L2 and Softmax for the top 5 accuracies. The proposed network works well with Softmax, with or without L2 normalization, as the CMC curves of the two techniques are similar. This suggests that the proposed regional LSTM learning model is general enough to create good gait-embedded vectors without normalization. The proposed regional LSTM with AdaCos and AMSoftmax perform significantly worse than the proposed with Softmax. This shows that scaling parameters and user-defined margins are not needed. It further implies that the proposed regional

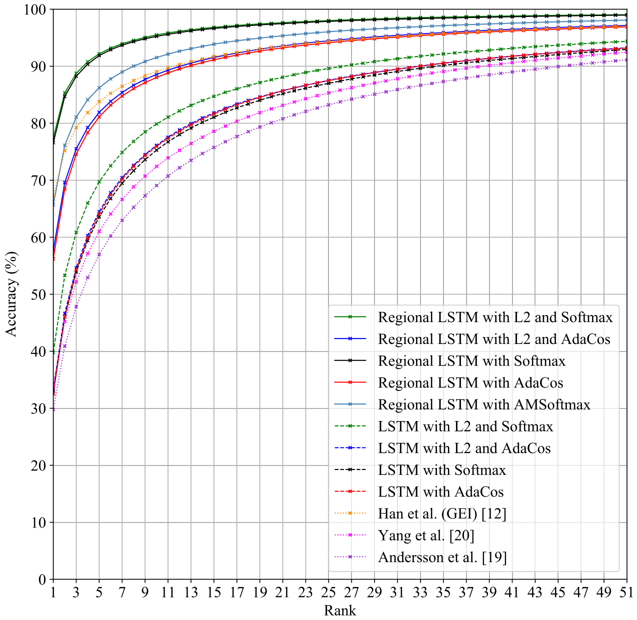

FIGURE 4. Cumulative Match Characteristic (CMC) curves of the top 51 ranks on imbalanced gallery.

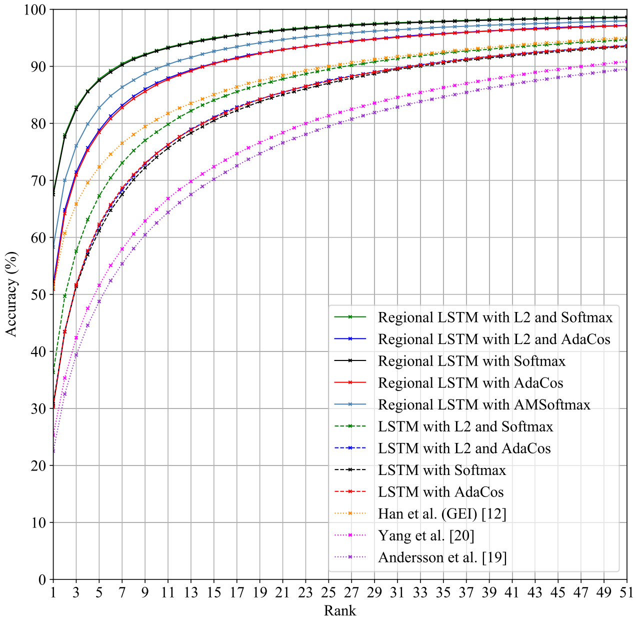

FIGURE 5. Cumulative Match Characteristic (CMC) curves of the top 51 ranks on balanced gallery.

TABLE 2. Performance of the proposed techniques compared to the existing techniques on imbalanced gallery.

Techniques	Top-k Accuracy (%) 1	2	3	4	5	Mean AUC (%) ROC	PR
Regional LSTM with L2 and Softmax	**77.4 ± 1.6**	**85.3 ± 1.2**	**88.8 ± 0.9**	**90.8 ± 0.7**	**92.2 ± 0.7**	86.9 ± 0.9	19.6 ± 1.5
Regional LSTM with L2 and AdaCos	57.7† ± 3.5	69.6† ± 2.8	75.5† ± 2.4	79.3† ± 2.0	81.9† ± 1.8	87.6 ± 0.9	18.8 ± 1.4
Regional LSTM with Softmax	76.6 ± 1.7	84.7 ± 1.3	88.2 ± 1.0	90.4 ± 0.9	91.9 ± 0.7	87.4 ± 1.0	**20.1 ± 1.5**
Regional LSTM with AdaCos	56.2† ± 3.1	68.5† ± 2.4	74.6† ± 2.0	78.4† ± 1.8	81.1† ± 1.6	**87.7 ± 1.0**	19.1 ± 1.6
Regional LSTM with AMSoftmax	65.7† ± 5.4	76.1† ± 4.3	81.1† ± 3.6	84.1† ± 3.1	86.2† ± 2.8	86.7† ± 1.1	18.6 ± 1.5
LSTM with L2 and Softmax	39.8† ± 1.8	53.3† ± 1.9	60.9† ± 1.8	66.0† ± 1.7	69.7† ± 1.6	85.2† ± 0.9	13.9† ± 1.0
LSTM with L2 and AdaCos	33.2† ± 2.0	46.6† ± 2.1	54.6† ± 2.0	60.3† ± 1.9	64.5† ± 1.8	84.5† ± 0.8	12.6† ± 0.8
LSTM with Softmax	32.9† ± 6.7	46.0† ± 8.0	53.9† ± 8.3	59.4† ± 8.3	63.6† ± 8.1	83.9† ± 2.6	12.1† ± 2.2
LSTM with AdaCos	32.5† ± 3.1	46.1† ± 3.3	54.2† ± 3.2	59.8† ± 3.1	64.1† ± 2.9	84.4† ± 1.3	12.5† ± 1.3
Han et al. (GEI) [12]	67.1† ± 1.7	75.2† ± 1.3	79.2† ± 1.2	81.9† ± 1.1	83.8† ± 1.0	61.6† ± 0.7	4.8† ± 0.2
Yang et al. [20]	34.0† ± 2.2	45.3† ± 2.1	52.2† ± 2.0	57.1† ± 2.0	61.0† ± 1.9	63.5† ± 2.1	4.1† ± 0.4
Andersson et al. [19]	29.8† ± 1.9	40.9† ± 2.0	47.8† ± 2.1	52.9† ± 2.0	57.0† ± 2.0	65.3† ± 2.3	4.5† ± 0.5

Note: techniques in boldface font show our proposed techniques; accuracy in boldface font shows the highest performance between techniques.
The AUC is area under curve.
† Indicates statistical significance from the maximum accuracy with $p = 0.05$.

TABLE 3. Performance of the proposed techniques compared to the existing techniques on balanced gallery.

Techniques	Top-k Accuracy (%) 1	2	3	4	5	Mean AUC (%) ROC	PR
Regional LSTM with L2 and Softmax	**67.8 ± 2.0**	**78.0 ± 1.5**	**82.8 ± 1.1**	**85.7 ± 1.0**	**87.8 ± 0.8**	86.6 ± 0.7	18.7 ± 1.2
Regional LSTM with L2 and AdaCos	51.8† ± 3.5	64.8† ± 2.7	71.5† ± 2.1	75.7† ± 1.7	78.8† ± 1.5	87.2 ± 0.6	17.8† ± 1.0
Regional LSTM with Softmax	67.5 ± 2.0	77.6 ± 1.6	82.4 ± 1.3	85.5 ± 1.1	87.5 ± 1.0	**87.2 ± 0.7**	**19.3 ± 1.2**
Regional LSTM with AdaCos	50.9† ± 2.6	64.2† ± 2.2	71.0† ± 1.8	75.3† ± 1.7	78.4† ± 1.5	87.2 ± 0.8	18.0† ± 1.3
Regional LSTM with AMSoftmax	58.3† ± 4.5	70.0† ± 3.7	76.1† ± 3.3	79.9† ± 2.8	82.7† ± 2.3	86.3† ± 0.8	17.6† ± 1.2
LSTM with L2 and Softmax	36.3† ± 1.9	49.7† ± 2.0	57.6† ± 2.0	63.1† ± 2.1	67.3† ± 2.0	85.3† ± 0.8	13.4† ± 0.9
LSTM with L2 and AdaCos	30.5† ± 1.9	43.5† ± 2.0	51.6† ± 1.8	57.4† ± 1.8	62.0† ± 1.5	84.2† ± 0.8	11.9† ± 0.7
LSTM with Softmax	30.8† ± 5.7	43.4† ± 6.9	51.4† ± 7.4	56.9† ± 7.5	61.2† ± 7.3	84.0† ± 2.6	11.6† ± 2.0
LSTM with AdaCos	30.3† ± 3.0	43.4† ± 3.4	51.7† ± 3.3	57.7† ± 3.2	62.3† ± 3.1	84.3† ± 1.0	11.9† ± 1.1
Han et al. (GEI) [12]	51.3† ± 1.9	60.7† ± 1.6	65.9† ± 1.5	69.6† ± 1.3	72.4† ± 1.2	58.2† ± 0.5	4.0† ± 0.2
Yang et al. [20]	25.3† ± 1.9	35.3† ± 2.2	42.4† ± 2.3	47.5† ± 2.4	51.6† ± 2.3	63.2† ± 1.8	3.8† ± 0.4
Andersson et al. [19]	22.5† ± 1.6	32.5† ± 1.8	39.4† ± 1.9	44.6† ± 1.8	48.8† ± 1.9	65.3† ± 1.8	4.1† ± 0.4

Note: techniques in boldface font show our proposed techniques; accuracy in boldface font shows the highest performance between techniques.
The AUC is area under curve.
† Indicates statistical significance from the maximum accuracy with $p = 0.05$.

LSTM learning model already creates gait-embedded vectors that effectively represent the gait identity of a person from only a 2-second walk.

Experimental results show that single LSTM with various loss functions provide significantly less accuracy than the regional LSTM models in all 5 ranks on the imbalanced gallery. Moreover, CMC curves of all single LSTM models are lower than the regional LSTM models throughout ranks 1 to 51. This illustrates that the regional LSTM learning model can extract a more unique identity of a person from gaits than just one single LSTM. This may be because each region of the body has its own rhythm when a person walks. For example, arm movements may have a different pattern than hip movements. The unique characteristics of regional movements may be lost by considering movement of the entire body.

Existing techniques, [12], [19], and [20] also perform significantly worse than the proposed regional LSTM techniques. As mentioned earlier, most gait recognition techniques are designed to work in supervised situations where subject identities are required. Consequently, most gait recognition techniques cannot be implemented in unsupervised situations. These three existing techniques were originally designed to be used in supervised situations. Moreover, these techniques were originally designed to handle walks that are much longer than 2 seconds. They may not be suitable in unsupervised 2-second walk situations. All three existing techniques, [12], [19], and [20], construct gait features from an entire walk, not sequentially. These gait features are single vectors representing the entire body as a whole. This means that they do not use the order of the movements in a walk or regional movements. This may lead to lower accuracies in all ranks of CMC accuracies, as shown in Fig. 4.

Interestingly, GEI (Han *et al.* [12]) provides higher accuracies than single LSTM techniques. Since GEI is a heat map of a body's silhouettes in a walk, unique movements of some regions of a body may be reflected in a GEI, which may be lost in a single GEI. This further supports that unique movements of regions of the body are crucial in gait recognition. However, GEI still performs significantly worse than the proposed regional LSTM techniques. This suggests that sequences of regional movements should also be considered.

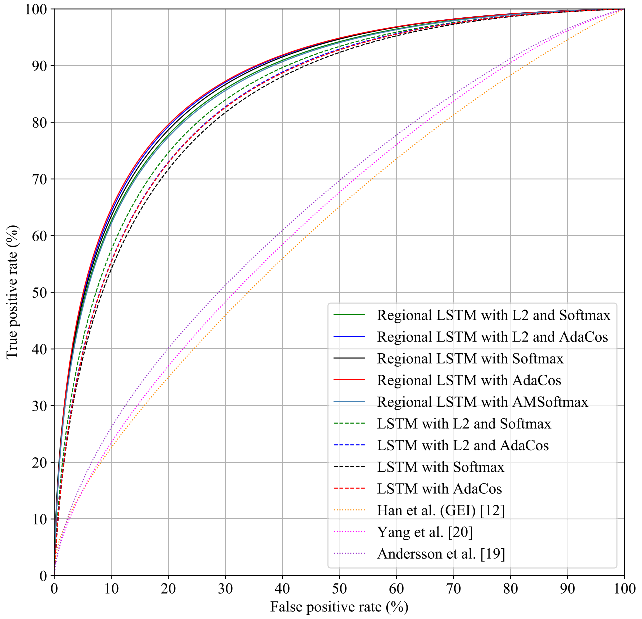

FIGURE 6. Receiver Operating Characteristic (ROC) curves on imbalanced gallery.

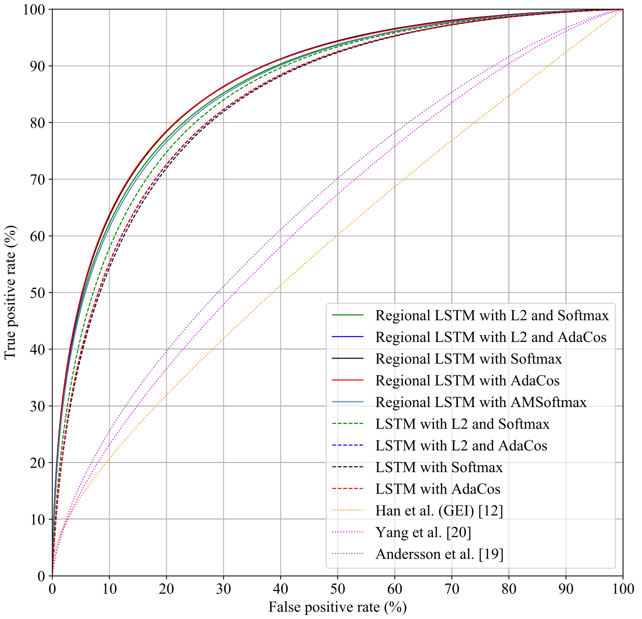

FIGURE 7. Receiver Operating Characteristic (ROC) curves on balanced gallery.

2) BALANCED GALLERY

Similar to the experimental results in the imbalanced gallery, the proposed regional LSTM with L2 and Softmax significantly outperforms other techniques, except the proposed regional LSTM with Softmax, as displayed in Table 2 and Fig. 4. The results confirm that the proposed regional LSTM with Softmax (with or without L2) has a high-ranking performance on both the balanced and the imbalanced galleries.

B. RECEIVER OPERATING CHARACTERISTICS (ROC) CURVES

Consider a manhunt scenario where authorities have a short poor-quality clip of a real criminal from one CCTV camera and would like to know where the real criminal is by searching for other clips from different CCTV cameras. A productive recognition and re-identification technique should be able to retrieve clips that mostly are the real criminal and only a small number or none are others. In the other words, these techniques should possess high separability ability. The Receiver Operating Characteristic (ROC) curves are measurements for this ability. ROC curves show the true positive rate vs the false positive rate of recognition and re-identification techniques. ROC curves of the proposed techniques and existing techniques, [12], [19], [20], on the imbalanced and balanced galleries are shown in Fig. 6 and Fig. 7, respectively.

1) IMBALANCED GALLERY

Experimental results on the imbalanced gallery show that ROC curves of all proposed regional LSTM techniques are higher than LSTM techniques and the existing techniques [12], [19], and [20]. From Table 2, the proposed regional LSTM with AdaCos obtains the highest area under the ROC curve. However, other regional LSTM techniques, except the regional LSTM with AMSoftmax, obtain high areas under the ROC curves that are not significantly different from the highest result. Consequently, the results indicate that the proposed regional LSTM models, except the regional LSTM with AMSoftmax, can separate gaits of one subject from others.

Areas under the ROC curves of single LSTM techniques are all above 80%, which are high but still significantly less than the proposed regional LSTM techniques. Areas under the ROC curves of existing techniques [12], [19], and [20], are around 60%. This suggests that the sequence of data and a separate pattern of regions of the body are needed, as used in the proposed regional LSTM techniques, for a productive recognition and re-identification technique.

2) BALANCED GALLERY

The proposed regional LSTM with Softmax obtains the highest area under the ROC curve on the balanced gallery but not significantly higher than the area under the ROC curve of the proposed regional LSTM with L2 and Softmax. The highest area under the ROC curve is higher than those of the single LSTM and existing techniques. This shows that the proposed regional LSTM techniques can distinguish the gaits of one subject from others significantly better than other techniques on both the balanced and the imbalanced galleries.

C. PRECISION-RECALL (PR) CURVES

Consider the same scenario as described at the beginning of Section IV-B. A competent recognition and re-identification technique should be able to retrieve almost all clips that belong to the real criminal. The Precision-Recall (PR) curve test is a measurement to asset the relevancy of recognition

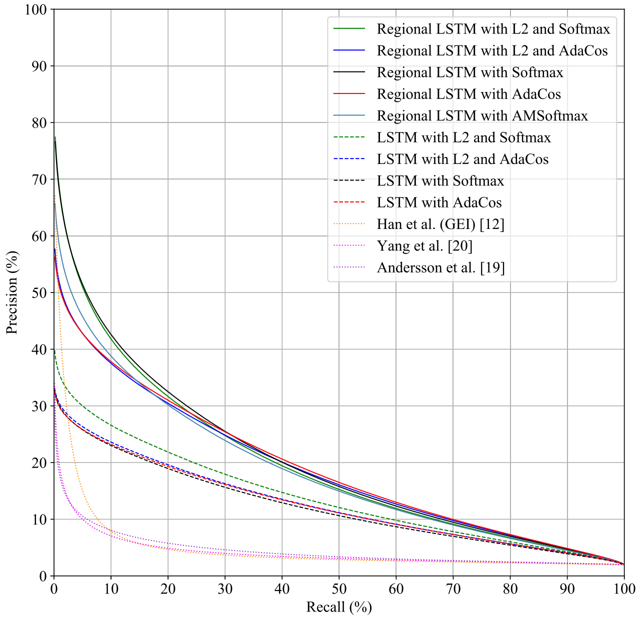

FIGURE 8. Precision-Recall (PR) curves on imbalanced gallery.

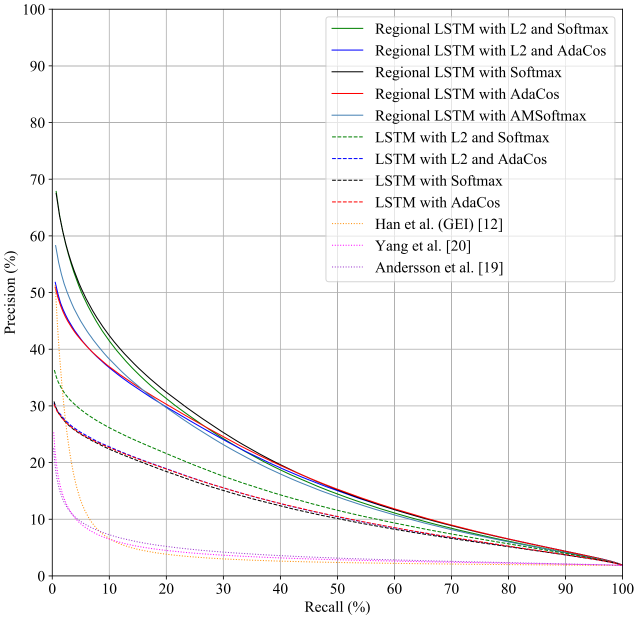

FIGURE 9. Precision-Recall (PR) curves on balanced gallery.

and re-identification techniques for this purpose. PR curves exhibit precision versus recall (true positive rate) of techniques. PR curves of the proposed techniques and existing techniques, [12], [19], [20], on the imbalanced and balanced galleries are shown in Fig. 8 and Fig. 9.

1) IMBALANCED GALLERY

Experimental results show that the area under the PR curve of the proposed regional LSTM with Softmax obtains the highest area, which is significantly higher than all single LSTM and existing techniques, but not significantly different from the rest of the regional LSTM techniques. Results indicate that the proposed regional LSTM techniques have significantly higher relevancy than the rest.

2) BALANCED GALLERY

On the balanced gallery, the area under the curve of the proposed regional LSTM with Softmax contains the highest area which is significantly higher than other techniques, except the proposed regional LSTM with L2 and Softmax. Unlike in the imbalanced gallery, the proposed regional LSTM with Softmax (with or without L2) has significantly higher relevancy than the existing techniques and other regional LSTM techniques. This implies that the LSTM with Softmax performs well under the PR curve test on both the imbalanced and balanced galleries.

D. ADVANTAGES OF GAIT RECOGNITION AND RE-IDENTIFICATION BASED ON REGIONAL LSTM

A person's gait consists of joints movements, where each joint has its own different pattern. For example, a way a person moves his or her head during a walk is different from the way his or her left shoulder moves. Moreover, the interaction between joint patterns is also necessary to develop a gait characteristic. Unlike most existing techniques where patterns of movements of the entire bodies are observed, the proposed regional LSTM gait recognition and re-identification is designed to capture and enhance unique joint patterns as well as their interactions. When gait features are created from movements of the entire bodies like in most existing techniques, some nuance but unique movements of some joints are diluted in the entire body movements. Since it is designed to focus on unique pattern of each region separately, the proposed method is able to enhance these subtle unique movements without getting lost in the movements of the entire body.

The proposed regional LSTM technique utilizes sequential gait data, unlike most existing techniques where orders of movement are often ignored e.g., some existing techniques use sums, averages, or standard deviations of gait data. Experimental results show that performances of existing techniques that ignore order of movements perform poorly in comparison to the proposed technique. This suggests that the sequence of movements are vital to identify a person.

The proposed technique is designed to handle the view point issue, whereas many existing techniques are often designed for fixed direction walks. In a real world situation, a person walks in all directions and often makes many turns. Gait recognition and re-identification techniques, like the proposed technique, that endure the view point issue are more suitable to be used in the real world applications.

Many existing techniques are supervised techniques where gaits of known persons are required. The proposed technique is an unsupervised technique where it can be used with or without gaits of known persons. This advantage allows the proposed technique to be used in a wider range of applications.

Since the proposed technique is designed to capture, and enhance subtle movements of separate regions of the body and how they interact with others, its applications are not limited to be used in people with normal gaits. It can be used with people with pathological gaits or abnormal gaits. In fact, joint movements of people with pathological gaits are very different from those of people with normal gaits. Consequently, gait embedded vectors of subjects with abnormal gaits, constructed from the proposed technique, would have greater distances from gait embedded vectors of people with normal gait. The proposed technique should be able to identify people with pathological gaits as well as those with normal gaits.

V. CONCLUSION

In this paper, we propose a new gait recognition and re-identification technique based on a proposed regional LSTM learning model. The proposed technique is designed to handle a short freestyle 2-second walk. It is created based on the idea that each region of the body has its own rhythmic movement during a walk and some movements of some regions may have more unique characteristics than others. A separate LSTM model is created to extract meaningful information from sequential data of each region of a body. In this process, unique characteristics of a region of the body are obtained sequentially with preserves the rhythm of the regional movement. Then, the proposed method combines the output of all 22 regional LSTM models to create a gait-embedded vector. In combining all 22 regions into one feature, the proposed model assigns weights to different regions. Hence all regions may not carry equal weight in the recognition and re-identification process. On both balanced and imbalanced datasets, experimental results show that the proposed regional LSTM learning model outperforms the existing techniques significantly in all three popular human recognition and re-identifiable tests: Cumulative Matching Characteristics (CMC) curves and top-k accuracy, Receiver Operating Characteristic (ROC) curves, and Precision-Recall (PR) curves. This indicates that the proposed regional LSTM learning model has a high-ranking performance (CMC test), can productively separate gaits of a subject from others (ROC test), and possesses an efficient relevancy ability (PR test). Since subjects in the gallery are not part of the training set, the experimental results indicate that the proposed regional LSTM learning model can be used effectively for human recognition and re-identification without subject labeling, unlike most gait recognition where subject labeling is required. This implies that the proposed regional LSTM technique is suitable for assisting authorities in tracking and re-identifying a person of interest, especially the identity of an unknown.

REFERENCES

[1] J. P. Singh, S. Jain, S. Arora, and U. P. Singh, "Vision-based gait recognition: A survey," *IEEE Access*, vol. 6, pp. 70497–70527, 2018.

[2] S. D. Matovski, S. M. Nixon, and N. J. Carter, *Gait Recognition*. Boston, MA, USA: Springer, 2014, pp. 309–318.

[3] J. E. Boyd and J. J. Little, *Biometric Gait Recognition*. Berlin, Germany: Springer, 2005, pp. 19–42.

[4] I. Rida, N. Almaadeed, and S. Almaadeed, "Robust gait recognition: A comprehensive survey," *IET Biometrics*, vol. 8, no. 1, pp. 14–28, Jan. 2019.

[5] P. Limcharoen, N. Khamsemanan, and C. Nattee, "View-independent gait recognition using joint replacement coordinates (JRCs) and convolutional neural network," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3430–3442, 2020.

[6] Y. Makihara, T. Tanoue, D. Muramatsu, Y. Yagi, S. Mori, Y. Utsumi, M. Iwamura, and K. Kise, "Individuality-preserving silhouette extraction for gait recognition," *IPSJ Trans. Comput. Vis. Appl*, vol. 7, pp. 74–78, Jul. 2015.

[7] 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. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017.

[8] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Computer Vision—ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Germany: Springer, 2006, pp. 151–163.

[9] J. Lu and Y.-P. Tan, "Uncorrelated discriminant simplex analysis for view-invariant gait signal computing," *Pattern Recognit. Lett.*, vol. 31, no. 5, pp. 382–393, 2010.

[10] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8126–8133.

[11] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2017.

[12] J. Man and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[13] M. Hofmann, S. M. Schmidt, A. N. Rajagopalan, and G. Rigoll, "Combined face and gait recognition using alpha matte preprocessing," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar. 2012, pp. 390–395.

[14] F. Battistone and A. Petrosino, "TGLSTM: A time based graph deep learning approach to gait recognition," *Pattern Recognit. Lett.*, vol. 126, pp. 132–138, Sep. 2019.

[15] Y. Feng, Y. Li, and J. Luo, "Learning effective gait features using LSTM," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 325–330.

[16] A. Cheewakidakarn, N. Khamsemanan, and C. Nattee, "View independent human identification by gait analysis using skeletal data and dynamic time warping," in *Proc. 14th Int. Symp. Adv. Intell. Syst. (ISIS)*, 2013, pp. 1–6.

[17] J. Preis, M. Kessel, M. Werner, and C. Linnhoff-Popien, "Gait recognition with Kinect," in *Proc. 1st Int. Workshop Kinect Pervasive Comput.*, Jun. 2012, pp. 1–4.

[18] R. M. Araujo, G. Graña, and V. Andersson, "Towards skeleton biometric identification using the Microsoft Kinect sensor," in *Proc. 28th Annu. ACM Symp. Appl. Comput. (SAC)*, New York, NY, USA, 2013, pp. 21–26.

[19] O. V. Andersson and R. M. D. Araújo, "Person identification using anthropometric and gait data from Kinect sensor," in *Proc. AAAI*, 2015, pp. 1–7.

[20] K. Yang, Y. Dou, S. Lv, F. Zhang, and Q. Lv, "Relative distance features for gait recognition with Kinect," *J. Vis. Commun. Image Represent.*, vol. 39, pp. 209–217, Aug. 2016.

[21] A. Ball, D. Rye, F. Ramos, and M. Velonaki, "Unsupervised clustering of people from 'skeleton' data," in *Proc. 7th Annu. ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, 2012, pp. 225–226.

[22] V. O. Andersson and R. M. Araujo, "Full body person identification using the Kinect sensor," in *Proc. IEEE 26th Int. Conf. Tools Artif. Intell.*, Nov. 2014, pp. 627–633.

[23] E. Gianaria, M. Grangetto, M. Lucenteforte, and N. Balossino, "Human classification using gait features," in *Biometric Authentication*, V. Cantoni, D. Dimov, and M. Tistarelli, Eds. Cham, Switzerland: Springer, 2014, pp. 16–27.

[24] E. Gianaria, N. Balossino, M. Grangetto, and M. Lucenteforte, "Gait characterization using dynamic skeleton acquisition," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2013, pp. 440–445.

[25] N. Jianwattanapaisarn, A. Cheewakidakarn, N. Khamsemanan, and C. Nattee, "Human identification using skeletal gait and silhouette data extracted by Microsoft Kinect," in *Proc. Joint 7th Int. Conf. Soft Comput. Intell. Syst. (SCIS), 15th Int. Symp. Adv. Intell. Syst. (ISIS)*, Dec. 2014, pp. 410–414.

[26] M. Ahmed, "Kinect-based human gait recognition using static and dynamic features," *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, pp. 425–431, Dec. 2016.

[27] C. Fengjiang, D. Muqing, and W. Cong, "Kinect-based gait recognition system design via deterministic learning," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, May 2017, pp. 5916–5921.

[28] F. Ahmed, P. P. Paul, and M. L. Gavrilova, "DTW-based kernel and rank-level fusion for 3D gait recognition using Kinect," *Vis. Comput.*, vol. 31, nos. 6–8, pp. 915–924, Jun. 2015.

[29] F. Ahmed, P. P. Paul, and M. Gavrilova, "Kinect-based gait recognition using sequence of the most relevant joint relative angles," *J. WSCG*, vol. 23, pp. 147–156, Jul. 2015.

[30] N. Khamsemanan, C. Nattee, and N. Jianwattanapaisarn, "Human identification from freestyle walks using posture-based gait feature," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 119–128, Jan. 2018.

[31] A. S. M. H. Bari and L. M. Gavrilova, "Multi-layer perceptron architecture for Kinect-based gait recognition," in *Proc. Adv. Comput. Graph. 36th Comput. Graph. Int. Conf. (CGI)*, in Lecture Notes in Computer Science, vol. 11542, M. L. Gavrilova, J. Chang, N. Magnenat-Thalmann, E. Hitzer, and H. Ishikawa, Eds., Calgary, AB, Canada. Cham, Switzerland: Springer, Jun. 2019, pp. 356–363.

[32] C. Nattee and N. Khamsemanan, "A deep neural network approach for model-based gait recognition," *Thai J. Math.*, vol. 17, no. 1, pp. 89–97, 2019.

[33] S. Choi, J. Kim, W. Kim, and C. Kim, "Skeleton-based gait recognition via robust frame-level matching," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2577–2592, Oct. 2019.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 815–823. [Online]. Available: https://ieeexplore.ieee.org/document/7298682

[35] R. Ranjan, C. D. Castillo, and R. Chellappa, "L_2-constrained softmax loss for discriminative face verification," Jun. 2017, *arXiv:1703.09507*. [Online]. Available: https://arxiv.org/abs/1703.09507

[36] Y. Hou and C. Chen, "Local reconstruction error of l_2 norm for discriminant feature extraction," in *Proc. 7th Int. Conf. Comput. Intell. Secur. (CIS)*, Y. Wang, Y.-M. Cheung, P. Guo, and Y. Wei, Eds., Sanya, China. Washington, DC, USA: IEEE Computer Society, Dec. 2011, pp. 1164–1168.

[37] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "NormFace: L_2 hypersphere embedding for face verification," in *Proc. 25th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2017, pp. 1041–1049.

[38] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. 33rd Int. Conf. Mach. Learn.*, New York, NY, USA, Jun. 2016, pp. 507–516. [Online]. Available: http://proceedings.mlr.press/v48/liud16.html

[39] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 6738–6746. [Online]. Available: https://ieeexplore.ieee.org/document/8100196

[40] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *CoRR*, vol. abs/1801.05599, pp. 926–930, Jul. 2018.

[41] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Dec. 2018, pp. 5265–5274. [Online]. Available: https://ieeexplore.ieee.org/document/8578650

[42] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proc. 33rd Int. Conf. Int. Conf. Mach. Learn.*, vol. 48, Dec. 2016, pp. 507–516.

[43] J. Deng, J. Guo, N. Xue, I. Cotsia, and S. P. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 2019, pp. 4685–4694. [Online]. Available: https://ieeexplore.ieee.org/document/8953658

[44] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jan. 2020, pp. 10815–10824. [Online]. Available: https://ieeexplore.ieee.org/document/8953896

[45] M. Wang and W. Deng, "Deep face recognition: A survey," *CoRR*, vol. abs/1804.06655, pp. 471–478, Oct. 2018.

PIYA LIMCHAROEN received the bachelor's and master's degrees in computer sciences from Sirindhorn International Institute of Technology, Thammasat University, Thailand, where he is currently pursuing the Ph.D. degree in computer science.

NIRATTAYA KHAMSEMANAN received the bachelor's degree *(cum laude)* in mathematics from Cornell University, Ithaca, NY, USA, and the master's and Ph.D. degrees in mathematics from the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. She is currently an Associate Professor with Sirindhorn International Institute of Technology, Thammasat University, Thailand.

CHOLWICH NATTEE received the bachelor's degree in computer engineering from Chulalongkorn University, Thailand, and the master's and D.Eng. degrees in computer science from Tokyo Institute of Technology, Japan. He is currently an Associate Professor with Sirindhorn International Institute of Technology, Thammasat University, Thailand.