future internet

MDPI

Article

Movement Analysis for Neurological and Musculoskeletal Disorders Using Graph Convolutional Neural Network

Ibsa K. Jalata [1,*], Thanh-Dat Truong [1], Jessica L. Allen [2], Han-Seok Seo [3] and Khoa Luu [1]

[1] Computer Vision and Image Understanding Laboratory, Department of Computer Science and Computer Engineering, University of Arkansas, Fayetteville, AR 72701, USA; tt032@uark.edu (T.-D.T.); khoaluu@uark.edu (K.L.)
[2] Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown, WV 26506, USA; jessica.allen@mail.wvu.edu
[3] Department of Food Science, University of Arkansas, Fayetteville, AR 72701, USA; hanseok@uark.edu
* Correspondence: ikjalata@uark.edu

Abstract: Using optical motion capture and wearable sensors is a common way to analyze impaired movement in individuals with neurological and musculoskeletal disorders. However, using optical motion sensors and wearable sensors is expensive and often requires highly trained professionals to identify specific impairments. In this work, we proposed a graph convolutional neural network that mimics the intuition of physical therapists to identify patient-specific impairments based on video of a patient. In addition, two modeling approaches are compared: a graph convolutional network applied solely on skeleton input data and a graph convolutional network accompanied with a 1-dimensional convolutional neural network (1D-CNN). Experiments on the dataset showed that the proposed method not only improves the correlation of the predicted gait measure with the ground truth value (speed = 0.791, gait deviation index (GDI) = 0.792) but also enables faster training with fewer parameters. In conclusion, the proposed method shows that the possibility of using video-based data to treat neurological and musculoskeletal disorders with acceptable accuracy instead of depending on the expensive and labor-intensive optical motion capture systems.

Keywords: cerebral palsy; graph convolutional neural network; deep learning; 1D-CNN; gait parameters

Citation: Jalata, I.K.; Truong, T.-D.; Allen, J.L.; Seo, H.-S.; Luu, K. Movement Analysis for Neurological and Musculoskeletal Disorders Using Graph Convolutional Neural Network. *Future Internet* **2021**, *13*, 194. https://doi.org/10.3390/fi13080194

Academic Editors: Kaushik Roy, Mustafa Atay and Ajita Rattani

Received: 13 June 2021
Accepted: 19 July 2021
Published: 28 July 2021

1. Introduction

Over 40 million people in the United States are diagnosed with movement disorders, such as Parkinson's disease, stroke, dementia, cerebral palsy, osteoarthritis, multiple sclerosis, etc. Impaired movement often leads to a reduced ability to perform activities of daily living, decreased quality of life, and substantial societal costs (e.g., costs of health care, social care services, productivity loss, etc.) [1,2]. For example, in 2014 the Centers for Disease Control in the US estimated the lifetime costs of caring for an individual with cerebral palsy are approximately USD 1.3 billion [3]. Gait analysis is a popular method for diagnosing movement impairments in these populations, where it can be used to inform rehabilitative treatment and quantify the progress of improvements throughout the rehabilitative process.

The current gold standard for quantitative movement analysis is using optical motion capture systems [4,5], which require sensors to be placed on a subject at specific locations on the body. The 3D positions of each marker can then be triangulated by image sensors placed around the room. Although these systems can provide very accurate measurement of movement, there are several limitations preventing its wide use in clinical settings, such as costly equipment that is confined to laboratory settings, and the time-consuming accurate placement of sensors. Wearable sensors, such as inertial-measurement-units, have gained some traction for clinic-based quantitative movement analysis [6,7] because they overcome some of these limitations; in particular, the main benefit is that these wearable systems

can be taken out of the lab to measure more naturalistic movements. However, wearable systems are still expensive and require time consuming and accurate sensor placement.

Recent advances in computer vision-based tracking of videos show promise for gait analysis occurring in natural environments without requiring expensive equipment or placement of sensors. Such video data are easier to capture from patients and inexpensive to process. The potential clinical utility of video-based pose-tracking has recently been demonstrated in studies on infants [8], healthy adults [9], older adults at risk of falls [10], and children with cerebral palsy [11]. With regards to gait analysis, these studies have attempted to measure either joint kinematics [11] or spatiotemporal parameters, such as step length and cadence [9], from a single video camera and have achieved moderate accuracy. This moderate accuracy can be attributed in part to the machine learning techniques that have been employed. The purpose of our study was to use more advanced deep learning techniques to achieve better accuracy of video-based gait analysis compared to the gold-standard optical motion capture.

In our perspective, two barriers have prevented wider use of video based recognition for human gait analysis. First, interest in video based recognition that uses pose estimation for keypoints extraction is relatively new and has been primarily confined to researchers in computer science and related fields. Second, there is expectation for a validation of the gait parameters as calculated by video based recognition approaches against gold standard method, i.e., optical motion capture systems.

The goals of this study were two-fold: (1) extract the keypoints from a video using OpenPose [12], then predict quantitative gait metrics commonly used in clinical gait analysis; and (2) propose a novel method based on graph convolutional neural network used for the prediction of quantitative gait metrics and also can be applied for classification tasks. The proposed method were trained on 1792 videos of 1026 unique patients with cerebral palsy disorder [13]. The clinical gait analysis method was used on the optical motion capture data to calculate the gait parameter values and used as a ground truth for our proposed network. The gait parameters that were used as metrics in this study were walking speed, cadence, gait deviation index, and knee flexion angle at maximum extension.

2. Related Work

With current technology, it is possible to capture skeleton data in real-time using depth sensors and pose estimation algorithms [12] with less resource and computational demand. For the realization of motion dynamics, skeleton data are the best choice since they are robust to illumination change, complex background, and scene variation. Conventional methods like hand-crafted approaches and deep-learning approaches are common ways to extract skeleton data from images or videos. The hand-crafted approach focus on capturing the dynamic motion of the joints, such as the relative position of the joints [14], the covariance matrix of the joint trajectories [15], and also the design of several view-invariant features. Among these features design, group sparsity-based class-specific dictionary coding [16], rotation and translations of body parts [15,17], and canonical view of transformed features [18] are the common ones. Other traditional methods [19–21] combine the information from different modalities, for instance, from depth information and skeleton data to further enhance the performance. However, these methods do not capture the features needed to predict the gait parameters like deep learning methods. Recently, a deep learning approach amassed a lot of success in many fields. The approach preferred to traditional machine learning methods because it outperforms with less complicated models and without requiring extensive feature engineering. Among the deep learning methods, recurrent neural network (RNN)-based methods [13,22–24] that are known for the application of temporal dynamic behavior, and CNN-based methods are used in cases of parameter sharing and sparse connectivity to reduce the number of parameters that need to be learned [25]. Additionally, the CNN model mapped anatomical key points to an outcome metric (e.g., cadence) [11]. To improve performance, a two-stream-based model [26] integrates CNN and RNN that operate on RGB images and coordinate vectors

of skeletons ordered in temporal form, respectively. However, both the single network and the combined two-stream model come short in understanding human movement since the spatial dependence between the correlated joints could not be captured by the methods.

The main idea of a graph neural network representing the complex relationship and inter-dependencies between the data in non-Euclidean domains that is classified under recurrent graph neural network (RecGNNs) was conceived in [27] and further explained in [28,29]. These RecGNNs methods use neighbor information iteratively to learn a target node. The training continues until a stable fixed point is attained. In general, the methods focus on learning node representations with recurrent neural architectures. Such methods demand higher computation power. As a consequence RecGNNs methods [30,31] are proposed to address such problem. Analogous to the convolutional neural network in images and videos, a graph convolutional network is proposed to generalize the operation of convolution from grid data to graph data. Here, in graph convolutional network (ConvGNN) the node representation is obtained by aggregating its neighbor features and its own features. Unlike RecGNNs methods, ConvGNN stack multiple graph convolutional layers and also pooling layers to extract high-level node and graph representation. ConvGNNs fall into two main streams, spectral-based approaches [32–35] and spatial-based approaches [36–38]. Although spectral-based approaches focus on removing noises from graph signals, spatial methods define graph convolution by information propagation, an idea inherited from RecGNNs. Recently, graph convolutional neural networks (GCNNs) [34] have been proposed to break the gap between spectral and spatial-based approaches. In addition to the flexibility, and simplicity, spatial-based methods are more efficient for graph-related data. Thus, we chose the approach in this paper. Initially, a graph convolutional neural network is proposed for spatial dependency. For applications like traffic forecasting, action recognition, CNN or RNN is used for temporal dependency alongside graph convolutional neural networks. Spatial-temporal graph neural network [39] address the time series prediction problem in traffic domain by applying graph convolutional neural network to both spatial and temporal dependency.

However, the method uses adjacency matrix only for spatial dimension. In the method, convectional convolutional neural network is used across temporal dimension that do not capture the features of skeleton dataset very well. Our method defines a single adjacency matrix that considers both temporal and spatial dimension. Moreover, the method gives a much faster training speed with fewer parameters.

Based on the landmark obtained by OpenPose [4], deep learning methods show promising results in gait analysis. Although OpenPose demonstrate higher location error of between 20 and 40 mm for laboratory based gait analysis comparing to the marker-based motion capture system, the method estimate a good landmark that can be used for further analysis [40]. CNN based method [11] predicts gait metrics that approach the theoretical limits for accuracy imposed by natural variability within the dataset prepared in Gillete Children Speciality Healthcare. However, still there is a room for improvement, where our method, i.e., graph convolutional neural network (GCNN) that fit the nature of the data improves the gait metrics.

The aim of this paper is producing gait metrics of video based system that mimics the metrics calculated from marker-based optical motion capture system. Beside using minimum squared error (MSE) during training, the application of information theory [41–45] namely correlation is used to show how the predicted value from our method is related to the ground truth value. For our dataset it turned out that the correlation values approach the maximum limit imposed by natural variability in the gait metrics.

3. The Proposed Method

3.1. Problem Formulation

Let $\mathbf{X} = q_1, \ldots, q_T$ denote a temporal sequence where each frame $q_t \in \mathbb{R}^{n_j \times 2}$ represents a human body pose at time t, with n_j number of joints in the skeleton, and each joint with 2 dimensions ((x, y) position on Cartesian plane). The OpenPose [12] method extracts the

joints from each video segment and we used the joints as input $\mathbf{X} \in \mathcal{X}$ for our proposed method after flattening it. Let $\mathbf{Y} \in \mathcal{Y}$ be the corresponding gait parameters of the input $\mathbf{X}$.

Let $\mathcal{F}$ be a non-linear function that employs the mapping from $\mathcal{X} \subseteq \mathbb{R}^{T \times n_j \times 2}$ to $\mathcal{Y} \subseteq \mathbb{R}^{1 \times 1}$, i.e., $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$. We parameterize the non-linear function $\mathcal{F}$ by a deep neural network with parameters θ_F. In general, given a dataset with N training samples, i.e., $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$, we will learn the network $\mathcal{F}(;\theta_F)$ so that can predict the gait parameter. In particular, the learning objective can be defined as follows:

$$\theta_F^* = \arg\min_{\theta_F} \mathbb{E}_{(\mathbf{X},\mathbf{Y})\in\mathcal{D}} \mathcal{L}(\mathcal{F}(\mathbf{X};\theta_F), \mathbf{Y}), \tag{1}$$

where $\mathcal{L}$ is the loss function between the predict gait parameter $\mathcal{F}(\mathbf{X};\theta_F)$ and the corresponding ground truth $\mathbf{Y}$.

3.2. Data Preprocessing

To analyze the gait parameter of patients, we extract skeleton anatomical data from the Gillete Children Speciality Healthcare dataset using OpenPose [5]. The extracted data have 25 key points for each frame. Here, each keypoint are taken as nodes and the x, y location of each point is represented as a feature of the node. To suit our model, we flattened the keypoints inputs for a segment of a video, $\mathbf{X}$. Despite Openpose [5] shows an efficient pose estimation method, some significant data are missing during extraction. Such missing data might contribute to the inaccurate prediction of gait metrics. We address the problem by replacing each missing joints with the mean of the feature's the probably extracted joints.

3.3. Graph Convolutional Neural Network

There are two approaches to define convolutional filters in a convolutional neural network: spectral-based graph convolution and spatial-based graph convolution. The spectral-based graph convolution has a solid mathematical foundation and it produced a good result in some applications. However, the method is not flexible to apply to many structures. The filter defined is domain-dependent. Additionally, the eigen decomposition in spectral-domain costs higher computational complexity. Like the typical convolutional method applied on videos, images, and sounds, a spatial convolutional method is implemented based on spatial relationships of the entity-nodes.

The skeleton of the body is represented as an undirected graph $G = \{V, E\}$ on a skeleton sequence with N joints and T frames featuring both intra-body and inter-frame connection. In this graph, the node set $V = \{v_i | i = 0, \ldots, N-1\}$ includes all the joints in a skeleton sequence. Instead of taking the spatial and temporal features as a separate entity, we joined them into a single dimension. Therefore, the set of joints V consists of all joints from intra-body and inter-frame connection. Thus, the total number of nodes, $N = n_j \times T$, where n_j is number of joints per frame and T is number of frames. Figure 1 shows the connectivity of joints in a frame and the connection of the same joint in consecutive frames.

In spatial and temporal graph convolution where the edge is defined on both spatial and temporal dimensions the edge, set E is composed of two subsets, the first subset depicts the intra-skeleton connection at each frame, and the second subset contains the inter-frame edges, which connect the same joints in consecutive frames. The intra-skeleton denoted as $E_S = \{v_{ti}, v_{tj} | (i, j) \in n_j\}$ and inter-frame edges denoted as $E_T = \{v_{ti}, v_{(t+1)i}\}$, where n_j is the set of joints. For two consecutive joints in skeleton body, if the joints are connected with a single bone, we set the value of the edge between them 1. For intra-body connectivity, we set the edge between the same joint in consecutive frames to 1. We set a small value $\delta = 0.01$ for joints that are not connected in intra-body or inter-frame connectivity. The connectivity of all joints in a given video is represented in adjacency matrix, $\mathbf{A}$ as it is shown in Equation (2).

$$\mathbf{A} = \begin{pmatrix} d_{0,0} & d_{0,1} & d_{0,2} & ... & d_{0,N-1} \\ d_{1,0} & d_{1,1} & d_{1,2} & ... & d_{1,N-1} \\ d_{2,0} & d_{2,1} & d_{2,2} & ... & d_{2,N-1} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ d_{N-1,0} & d_{N-1,1} & d_{N-1,2} & ... & d_{N-1,N-1} \end{pmatrix} \quad (2)$$

where $d_{i,j}$ is the edge between two joints. As it is stated earlier, if there is no edge between two joints then $d_{i,j}$ set to $\delta = 0.01$. In Equation (2), the row of the adjacency matrix shows the connection of joint 0 with all joints in intra-skeleton and inter-frame. Let the number of frames is M and each frame has 4 joints. The total number of joints would be, $N = MX4$. Thus, in Equation (2), $d_{i,j}$ means the connection between the first joints in the first frame and the second joint which is located in the same frame. If there is an edge between the joints the value of $d_{i,j}$ set to 1, otherwise it is set to 0. The toy example in Figure 2 describe for node, $N = 16$. Each rows shows how node i is related to all of the nodes in the input. Since we organized the spatial, temporal data and the adjacency matrix as a 2D dimension, we can directly use the graph convolutional definition from [34] into our problem.

$$\mathbf{f}_{out} = \mathbf{D}^{-1/2}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-1/2}\mathbf{f}_{in}\mathbf{W} \quad (3)$$

where, $\mathbf{I}$ is the identity matrix representing self connections and $\mathbf{D}$ is the diagonal node degree matrix used for normalization given as the summation of adjacency matrix across the column, i.e., $\mathbf{D} = \sum_j \mathbf{A}_{i,j} \times \mathbf{f}_{in}$ is the feature with $\mathbb{R}^{N \times d}$ dimension and before applying the first graph convolution $\mathbf{f}_{in}$ considered as the spatial location of each node with $\mathbb{R}^{N \times 2}$ dimension.

Figure 1. The keypoints denoted as blue dots in first frame and green dots in the following frames are used as input to the proposed graph convolutional neural network. The dots denote the body joints of the subject. The solid line shows the natural connection in intra-body. The dot line connects the inter-frame edges connects the same joints in consecutive frames.

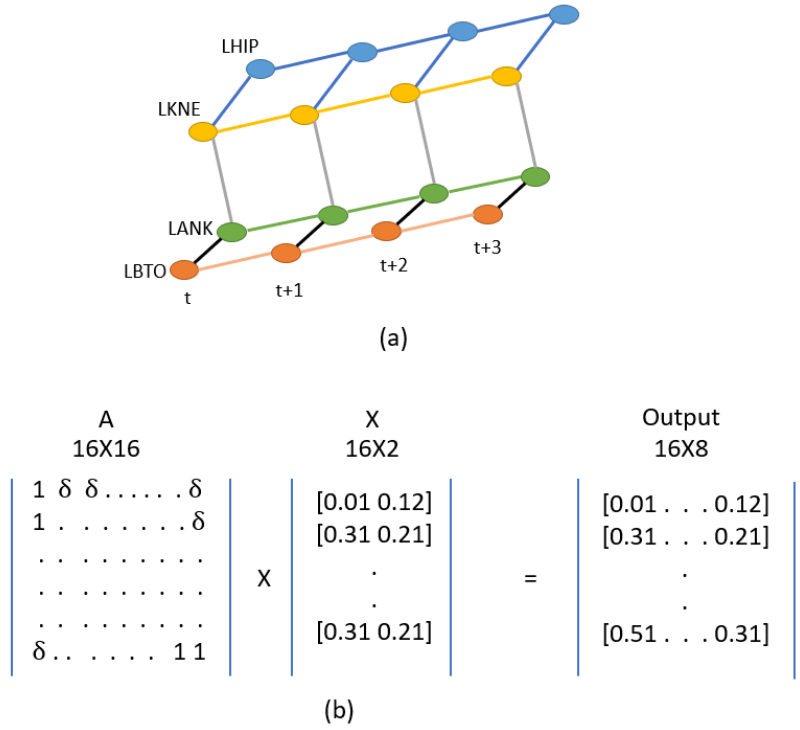

Figure 2. Illustration of (**a**) proposed method using an example of $N = 16$, where N is the total number of keypoints in 4 frames. Each point at t shows the joints of left skeleton. The left hip (LHIP) at t is connected to the left knee (LKNE) and LHIP at $t + 1$. (**b**) For the first graph convolutional layer the (x, y) position of each joints is taken as a feature. The adjacency matrix **A** describes the relation of each joints with each other. The LHIP joint in the frame t is connected with the LANK in the same frame t and with LHIP in frame $t + 1$. Hence, for LHIP in frame t, 1 is assigned for joints that have direct connection. Basically, The ouput feature is computed from matrix multiplication of the input feature X and The adjacency matrix **A**. To avoid the impact of zero value in matrix multiplication we set a very small value $\delta = 0.01$ for the joints that are not connected.

For readers' understanding, Figure 2 depicts a toy example of the proposed graph convolutional method. As it is explained earlier, among 25 keypoints, 8 keypoints which are directly related to gait measurements are selected. The selected key points are divided into left and right key points to process further. The (x, y) location of left key points (left-hip (LHIP)), left-knee (LKNE), left-ankle (LANK), and left-big-toe (LBTO) for t frames used as input feature for the first graph convolutional layer. In our model, we have used two graph convolutional layer; the first layer takes the location of each keypoints in a video segment as input and produce 8 features. The output of the first layer fed into second layer and produces features with 16 channel which is fed in to conventional convolutional layer for down sampling and finally the output would be predicted. Figure 3 further illustrates the pipeline of our method from input to output. This method can be applied to the classification problem. The only modification expected would be adding softmax after the fully connected layer. It is worth noting that, although we achieve a good result, using only key points extracted using OpenPose, we attained a better result when we comprise a hand-engineered time series data crafted from the relationship of the joints.

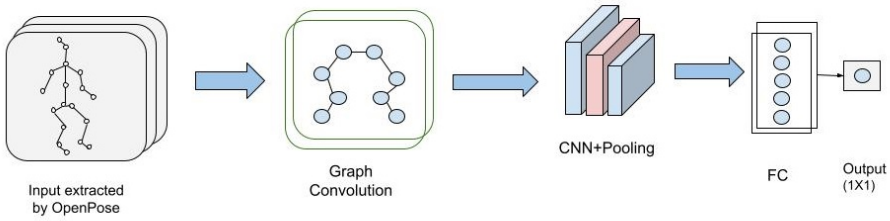

Figure 3. The proposed graph convolutional neural network (GCNN).

In addition to skeleton body joints, using derived time series data as input improves the overall performance of the network model. The first derived time series is computed as the angle between the vector from the knee to the hip and the vector from the knee to the ankle. The second time series was the difference between the x-coordinates of the left and right ankles. Our second model takes both types of input, i.e., skeleton joints and derived time series data, and predicts the gait parameters. As it is shown in Figure 4 a 1D convolutional neural network is applied on the derived time series data, and the output from this layer concatenated with the output from graph convolutional neural network. Then, the combined features passed through another 1D CNN and average pooling to learn more features and attain translation invariance. Specifically, average pooling computes the average of features of the skeleton joint for the segment of the video.

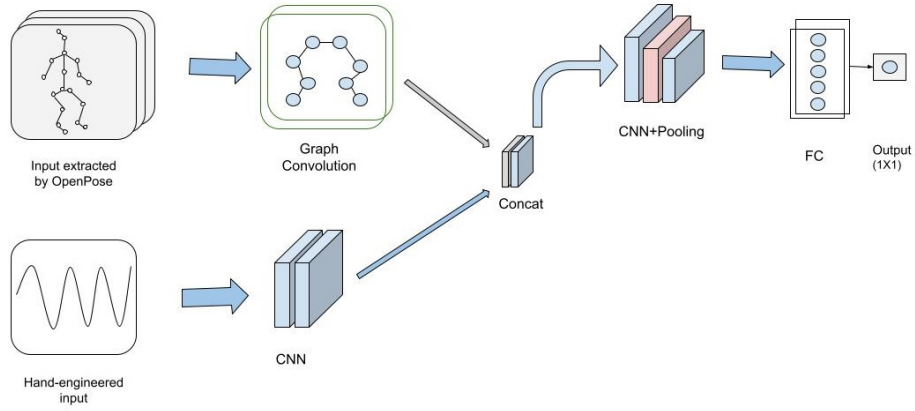

Figure 4. Graph convolutional neural network (GCNN) with hand-engineered data processed with convolutional neural network (CNN).

Since our proposed method is categorized under regression problem, we have used the most common loss function mean squared error (MSE). In Equation (4), $\mathbf{Y}_i$ is the ground truth for gait metrics evaluated from optical motion capture sensors. Based on reflective markers placed on patients the high-frequency cameras and motion capture software tracked the 3D positions and the 3D joint kinematics computed using the inverse kinematics [11]. Then, the time-series data of 3D-joint kinematics is analyzed and the gait metrics that are used as ground truth, $\mathbf{Y}_i$ is computed. $\mathcal{F}(\mathbf{X};\theta_F)$ is the predicted gait parameters by our model from video inputs. Finally, the loss function is formulated as follows:

$$\mathcal{L}(\mathcal{F}(\mathbf{X};\theta_F),\mathbf{Y}) = \frac{1}{n}\sum_{n=1}^{n}(\mathbf{Y}_i - \mathcal{F}(\mathbf{X}_i;\theta_F))^2 \quad (4)$$

4. Experiments

4.1. Dataset

We have used a dataset from Gillette Children's Speciality Healthcare collected from 1994 to 2015 [5]. The dataset has 1792 videos of 1026 unique patients with the cerebral palsy condition. In the dataset, the patient's average age was 11 years with the standard deviation (SD) of 5.9. Average height and mass were 133 cm (SD = 22) and 34 kg (SD = 17), respectively. The ground truth metrics were computed from optical motion capture data. As it is described in [5], for each patient, reflective markers are placed on anatomical landmarks. Then, the optical motion capture system incorporated with tracking software captured the positions of the markers as the patients moved in controlled space.

At last, engineers post-processed these data and computed gait metrics that were used as a ground-truth. The video-based data used in our model for training were collected with exact setting but at a different time to the ground truth. The skeleton data are extracted from video-based data using the publicly available OpenPose [12] toolbox. The toolbox gave 2D coordinates (x, y) with a confidence score of C for 25 joints.

4.2. Performance Metrics

The measurement gait metrics used in these methods are common in many neurological and musculoskeletal disorders. These gait metrics that were used in our model were walking speed, cadence, knee flexion angle at maximum extension, and GDI. For performance criteria, we have used the correlation coefficients to compare the ground truth values that are prepared from optical motion sensor values. During training, MSE loss function with adaptive moment estimation (Adam) was used as optimization. After choosing the best parameter for each metric on validation test we took 300 samples from test data to examine how the predicted value from our model could be related to the label using correlation coefficients.

4.3. Experiment Settings

In this work, training and testing were implemented on a machine with Linux cluster CPU: Intel(R) Core i7-8700K CPU @ 3.7 $GHz \times 12$, and GPU: NVIDIA GeForce GTX 1080. The network was trained in a fully-supervised way with $L2$ loss function and using adaptive moment estimation (Adam) as the optimization method. We trained both models for a maximum of 100 epochs with a learning rate of 0.015 and early stopping with a window size of 10, i.e., we stopped training if the validation loss could not decrease for 10 consecutive epochs. To avoid early stoppage of training, we decrease the learning by a factor of 10 every 20,000 iterations. We have allocated 60% of the data for training, 20% for validation, and the rest are allocated for testing.

4.4. Data Normalization

From each frame in a video, 2D image-plane coordinates of 25 keypoints with the confidence of individual keypoints were extracted by OpenPose toolkit. From each detected person in a frame the given points were the x, y coordinates, in pixels, of the centers of the torso, nose, and pelvis, and centers of the left and right shoulders, elbows, hands, hips, knees, ankles, heels, first and fifth toes, ears, and eyes [12]. The toolkit missed detecting few people from the frame. We removed 1443 such cases from using for training. For some cases only few of the skeleton joints are missed. For such cases, we used linear interpolation to fill the missed points.

Some of the input data were noisy, so they might not give an expected result. For example, the x-coordinate of the left ear and few other time-series data were noisy and contributed undesired results. To mitigate the effects of these noisy data, we normalized the image-plane coordinates of knees, ankles, hips, big toes, projected angles of the ankle and knee flexion, the distance between the first toe and ankle, and the distance between the left ankle and right ankle [13]. In addition, using window slicing more time-series data were generated, and using an augmented dataset enabled avoiding variation in each starting frame.

4.5. Result

Figure 5 depicts the comparison of graph convolutional neural network (GCNN), CNN, ridge regression (RR), and random forest (RF). In the proposed method, we have used mean squared error (MSE) to train our model. MSE gives us the measure of how far the predictions were from actual output but we do not know whether we are under predicting the data or over predicting the data. Therefore, we have used additional metrics, also known as correlation, to evaluate the outcome. Here, the term correlation was used as how the predicted gait parameters from video inputs related to the gait parameters calculated from laboratory-based motion capture sensors. We have taken 300 gait metrics samples that were predicted by our model based on the test data and we compared these samples with the ground truth and examined how they related with each other. Thus, the correlation between the gait metrics speed from our model and the ground truth was 0.791 (0.742–0.853). For GDI the correlation was 0.792 (0.710–0.822). For cadence, knee maximum flexion our proposed method predicted a similar results with CNN [11]. During inference, our system takes 0.2 s to predict the gait parameter of the patient.

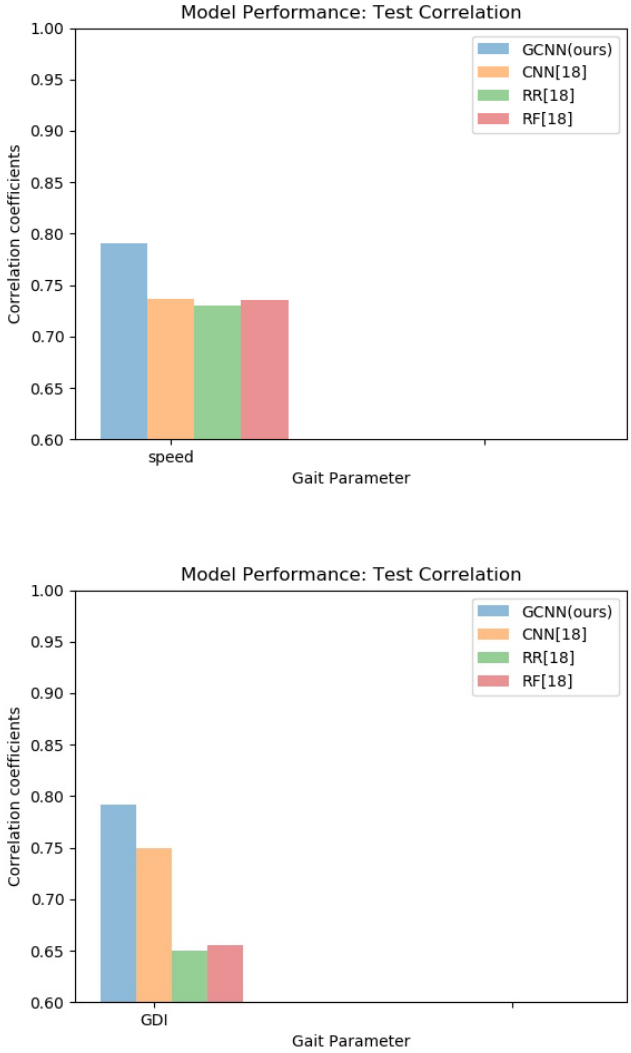

Figure 5. Comparison prediction accuracy of the proposed graph convolutional network with the previous methods on speed and gait deviation index (GDI) gait parameters. The three methods our proposed graph convolutional neural network (GCNN) compared are: convolutional neural network (CNN), ridge regression (RR), and random forest (RF). After training our model on video inputs, we have evaluated on test data. The correlation values shows how the predicted metric values are related with the ground truth that are prepared from optical motion sensor values.

Table 1 shows the detailed architecture of the proposed model (Figure 4). The 2D joints extracted from a video using OpenPose [12] were fed into the first part of the model. In addition to the keypoints extracted using OpenPose, we comprised a hand-engineered time serious derived from to get an optimal result. Concatenating the time-series data (computed from the relation of the skeleton joints) that is passed through 1D-CNN with GCNN increases the prediction accuracy. Table 2 shows the comparison of these two approaches in terms of correlation coefficient with the ground truth. We have the first two graph convolutional layers in part i. The graph convolutional layers used to compute 8 and 16 dimension feature maps, respectively. Part ii of the proposed method has two convolutional layers with 7 kernel size each. In part iii, i.e., after the concatenation of the joint features from graph convolutional layer and the time series features from the convolutional layer, we have used a 1D convolutional layer, followed by maxpooling and another 1D convolutional layer with size 8 kernels . All the layers are followed by batch norm and Relu activation.

Table 1. Layer descriptions of the proposed method: the first part of the graph convolutional network is composed of two graph convolutional layers, the second part is 1D convolutional for hand engineered input data and the third part of the layers comprise 1D convolutional, maxpooling, and fully connected layer for the concatenated input from part i and ii.

Part	Layer Type	Layer	Number of Unit	Kernel Size	Dropout
i	Graph Convolutional	1	8	-	0
	Graph Convolutional	1	16	-	0
ii	Convolutional	1	8	7	0
	Convolutional	1	16	7	0
	Concatenation i and ii	-	16		-
iii	Convolutional	1	8	8	0.5
	Maxpooling	2	16	3	0.5
	Convolutional	2	8	8	0.5
	Flatten	-	16	-	0.5
	FC(Fully Connected NN)	1	96	-	0.5
	FC(Fully Connected NN)	1	20	-	0.5
	Output	-	1	-	-

Table 2. Comparison of our proposed graph convolutional neural network (GCNN) depicted in Figures 3 and 4. Here, we observe the method that takes the hand crafted inputs (Figure 4) has a better result.

	Correlation for only GCNN	Correlation for GCNN
Walking speed (m/s)	0.651 (0.618–0.694)	0.791 (0.742–0.853)
Gait Deviation Index	0.580 (0.542–0.624)	0.792 (0.750–0.822)
Cadence (strides/s)	0.632 (0.584–0.653)	0.785 (0.756–0.802)
Knee flexion (degrees)	0.681 (0.643–0.741)	0.832 (0.752-0.867)

5. Conclusions

In this paper, we proposed a graph convolutional neural network that can capture the characteristics and relationships of skeleton input data in the spatial and temporal dimensions. Our method takes advantage of the structural information possessed by the input skeleton human pose which is extracted using OpenPose. In the method, spatial and temporal features of the input are represented in a single adjacency matrix that helps to apply graph convolution in both dimensions. As a result, the method captures the main features of the motion of the patient that contribute to predicting the gait parameter. Our approach has experimented on the cerebral palsy disorder and it outperformed the

state-of-the-art method on the dataset that was processed by Gillette Speciality Healthcare. Our method predicted clinically apparent motion metrics from an ordinary video of patients with the cerebral palsy disorder. The method helps the clinicians to address the symptoms of neurological and musculoskeletal disorders without placing reflective markers on patients' anatomical landmarks which is very expensive and takes a lot of effort and time to diagnosis the patients. In addition, the method achieved the result with fewer parameters, faster training, and earlier convergence. In future work, we will apply the proposed method for other types of musculoskeletal and neurological disorders.

Author Contributions: I.K.J. and T.-D.T. proposed conceptualization, I.K.J. performed the experiment and wrote the first draft. K.L., J.L.A. and H.-S.S. assisted with revisions and improvements. The work was performed under the supervision and guidance of K.L. and J.L.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: As described in this paper, the data prepared by Gillette Speciality [5] Healthcare is processed by using the OpenPose algorithm [12]. Due to privacy of the patients, the video data are not publicly available. However, the preprocessed data are publicly available at [11].

Acknowledgments: The statements made herein are solely the responsibility of the authors. The authors would like to thank the University of Arkansas and Neuromechanics of Mobility Lab, West Virginia University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mutikainen, S.; Rantanen, T.; Alén, M.; Kauppinen, M.; Karjalainen, J.; Kaprio, J.; Kujala, U.M. Walking Ability and All-Cause Mortality in Older Women. *Int. J. Sports Med.* **2011**, *32*, 216–222. [CrossRef] [PubMed]
2. Ostir, G.V.; Berges, I.M.; Kuo, Y.-F.; Goodwin, J.S.; Fisher, S.R.; Guralnik, J.M. Mobility Activity and Its Value as a Prognostic Indicator of Survival in Hospitalized Older Adults. *J. Am. Geriatr.* **2013**, *61*, 551–557. [CrossRef] [PubMed]
3. Mendez, M.F. Early-onset Alzheimer's disease: Nonamnestic subtypes and type 2 AD. *Arch. Med. Res.* **2012**, *43*, 677–685. [CrossRef] [PubMed]
4. Lee, L.; Grimson, W.E.L. Gait analysis for recognition and classification. In Proceedings of the Fifth IEEE International Conference on Automatic Face Gesture Recognition, Washington, DC, USA, 21 May 2002.
5. Schwartz, M.H.; Trost, J.P.; Wervey, R.A. Measurement and management of errors in quantitative gait data. *Gait Posture* **2002**, *20*, 196–203. [CrossRef] [PubMed]
6. Horak, F.; King, L.; Mancini, M. Role of Body-Worn Movement Monitor Technology for Balance and Gait Rehabilitation. *Phys. Ther.* **2015**, *95*, 461–470. [CrossRef] [PubMed]
7. Al-Amri, M.; Nicholas, K.; Button, K.; Sparkes, V.; Sheeran, L.; Davies, J. Inertial Measurement Units for Clinical Movement Analysis. *Natl. Libr. Med.* **2018**, *18*. [CrossRef]
8. Chambers, C.; Prosser, L.; Johnson, M.J.; Kording, K.P. Computer vision to automatically assess infant neuromotor risk. *IEEE Trans. Neural. Syst. Rehabil. Eng.* **2019**, *28*. [CrossRef]
9. Stenum, J.; Cristina, R.; Ryan, T. Roemmich. Two-dimensional video-based analysis of human gait using pose estimation. *PLoS Comput. Biol.* **2020**, *17*, e1008935.
10. Huang, Z.; Liu, Y.; Fang, Y.; Horn, B.K. Video-based fall detection for seniors with human pose estimation. In Proceedings of the 4th International Conference on Universal Village (UV), Boston, MA, USA, 21–24 October 2018.
11. Kidzinski, L.; Yang, B.; Hicks, J.L.; Rajagopal, A.; Delp, S.L.; Schwartz, M.H. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nat. Commun.* **2020**, *11*, 4054. [CrossRef] [PubMed]
12. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
13. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently recurrent neural network (indrnn): Building a longer and deeper RNN. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5457–5466.
14. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
15. Evangelidis, G.; Singh, G.; Horaud, R. Skeletal quads: Human action recognition using joint quadruples. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 4513–4518.

16. Luo, J.; Wang, W.; Qi, H. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 1809–1816.
17. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human action recognition by representing 3d skeletons as points in a lie group. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 588–595.
18. Rahmani, H.; Mian, A. Learning a non-linear knowledge transfer model for crossview action recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2458–2466.
19. Hu, J.F.; Zheng, W.S.; Lai, J.; Zhang, J. Jointly learning heterogeneous features for RGB-D activity recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5344–5352.
20. Ohn-Bar, E.; Trivedi, M. Joint angles similarities and HOG2 for action recognition. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 465–470.
21. Rahmani, H.; Mahmood, A.; Huynh, D.Q.; Mian, A. Real time action recognition using histograms of depth gradients and random decision forests. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, 24–26 March 2014; pp. 626–633.
22. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
23. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4263–4270.
24. Zamir, A.R.; Wu, T.L.; Sun, L.; Shen, W.B.; Shi, B.E.; Malik, J.; Savarese, S. Feedback networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1308–1317.
25. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2017**, *521*, 436–444. [CrossRef] [PubMed]
26. Zhao, R.; Ali, H.; Van der Smagt, P. Two-stream RNN/CNN for action recognition in 3D videos. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 4260–4267.
27. Gori, M.; Monfardini, G.; Scarselli, F. A new model for learning in graph domains. In Proceedings of the 2005 IEEE International Joint Conference on Neural Networks, Montreal, QC, Canada, 31 July–4 August 2005; pp. 729–734.
28. Scarselli, F.; Gori, M.; Tsoi, A.C.; Hagenbuchner, M.; Monfardini, G. The graph neural network model. *IEEE Trans. Neural Netw.* **2009**, *20*, 61–80. [CrossRef] [PubMed]
29. Gallicchio, C.; Micheli, A. Graph echo state networks. In Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 18–23 July 2010; pp. 1–8.
30. Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R. Gated graph sequence neural networks. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
31. Dai, H.; Kozareva, Z.; Dai, B.; Smola, A.; Song, L. Learning steadystates of iterative algorithms over graphs. In Proceedings of the International Conference on Machine Learning, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018.
32. Henaff, M.; Bruna, J.; LeCun, Y. Deep convolutional networks on graph-structured data. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015.
33. Defferrard, M.; Bresson, X.; Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5 December 2016; pp. 3844–3852.
34. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the 5th International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
35. Levie, R.; Monti, F.; Bresson, X.; Bronstein, M.M. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Trans. Signal Process.* **2017**, *67*, 97–109. [CrossRef]
36. Atwood, J.; Towsley, D. Diffusion-convolutional neural networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5 December 2016; pp. 1993–2001.
37. Niepert, M.; Ahmed, M.; Kutzkov, K. Learning convolutional neural networks for graphs. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 2014–2023.
38. Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural message passing for quantum chemistry. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 1263–1272.
39. Yu, B.; Yin, H.; Zhu, Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. *arXiv* **2018**, arXiv:1709.04875.
40. Nakano, N.; Sakura, T.; Ueda, K.; Omura, L.; Kimura, A.; Iino, Y.; Fukashiro, S.; Yoshioka, S. Evaluation of 3D Markerless Motion Capture Accuracy Using OpenPose with Multiple Video Cameras. *Front. Sport. Act. Living* **2020**, *2*, 50. [CrossRef] [PubMed]
41. Pregowska, A.; Kaplan, E.; Szczepanski, J. How Far can Neural Correlations Reduce Uncertainty? Comparison of Information Transmission Rates for Markov and Bernoulli Processes. *Int. J. Neural. Syst.* **2019**, *29*, 1950003. [CrossRef] [PubMed]
42. Brette, R. Philosophy of the spike: Rate-based versus spike-based theories of the brain. *Front. Syst. Neurosci.* **2015**, *9*, 151. [CrossRef] [PubMed]

43. Crumiller, M.; Knight, B.; Kaplan, E. The measurement of information transmitted by a neural population: Promises and challenges. *Entropy* **2013**, *15*, 3507–3527. [CrossRef]
44. Pregowska, A.; Szczepanski, J.; Wajnryb, E. Temporal code versus rate code for binary Information Sources. *Neurocomputing* **2016**, *216*, 756–762. [CrossRef]
45. Lorenzo, P.M.D.; Chen, J.Y.; Victor, J.D. Quality time: Representation of a multidimensional sensory domain through temporal coding. *J. Neurosci.* **2009**, *29*, 9227–9238. [CrossRef] [PubMed]

Short Biography of Authors

Ibsa Jalata is a Ph.D. student at the University of Arkansas, in Arkansas. He received his M.Sc. degree from Chonbuk National University in South Korea where he had been working as research assistance in a Media and Communication lab. He completed his B.Sc. in Addis Ababa University in Ethiopia. He has been working as a Lecturer in Adama Science and Technology University, Ethiopia. His research interests include activity recognition, action localization, computer vision, face detection and recognition, and biomedical image processing.

Thanh-Dat Truong is currently a Ph.D. student at the Department of Computer Science and Computer Engineering of the University of Arkansas. He received his B.Sc. degree in Computer Science from the Honors Program, Faculty of Information Technology, University of Science, VNU in 2019. He was a research intern at Coordinated Lab Science, at the University of Illinois, at Urbana-Champaign, where he worked on generative adversarial networks in 2018. When Thanh-Dat Truong was an undergraduate student, he worked as a research assistant at the Artificial Intelligence Lab at the University of Science, VNU. Thanh-Dat Truong's research interests include face recognition, face detection, domain adaptation, deep generative models, and adversarial learning. His papers appear at top tier conferences, such as The IEEE Computer Vision and Pattern Recognition, Canadian Conference on Computer and Robot Vision, and ACM International Conference on Multimedia Retrieval. He is also a reviewer of top-tier journals and conferences including IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI), Journal of Computers Environment and Urban Systems, Workshop on Applications of Computer Vision (WACV), International Conference on Pattern Recognition (ICPR).

Dr. Khoa Luu is currently an Assistant Professor and the Director of Computer Vision and Image Understanding (CVIU) Lab in Department of Computer Science and Computer Engineering at University of Arkansas. He was the Research Project Director in Cylab Biometrics Center at Carnegie Mellon University (CMU), USA. He has received four patents and two best paper awards, and co-authored 60+ papers in conferences and journals. He was a vice chair of Montreal Chapter IEEE SMCS in Canada from September 2009 to March 2011. His research interests focus on various topics, including biometrics, image processing, computer vision, machine learning, multifactor analysis, and compressed sensing. He is currently a reviewer for several top-tier conferences and journals, such as: CVPR, ICCV, ECCV, NIPS, TPAMI, TIP, JPR, SP, etc.

Dr. Jessica L Allen is an assistant professor in Biomedical Engineering at West Virginia University. She received her B.S. degree in Mechanical Engineering from the University of Florida and Ph.D. in Mechanical Engineering from the University of Texas at Austin. She also completed a post-doctoral fellowship in the Neural Control of Movement at Emory University. Her research interests include identifying interactions between the neural, muscular, and skeletal systems that can inform interventional decisions to improve overall mobility during daily life.

Dr. Han-Seok Seo is an Associate Professor and Director of the University of Arkansas Sensory Science Center in the Department of Food Science at the University of Arkansas, Fayetteville. Dr. Seo is a creative sensory scientist who combines multidisciplinary backgrounds and skills in order to contribute to improved quality of life through healthy and happy eating behavior. His research interests include identifying mechanisms of multisensory interaction and integration with a focus on chemosensory cues, developing methods to improve eating quality, creating novel methodology of sensory evaluation, and investigating impacts of sensory disorders on eating quality. He holds two doctoral degrees, a Ph.D. in Food and Nutrition and a Doctor of Medical Science in Otorhinolaryngology from Seoul National University (Seoul, Korea) and the Technical University of Dresden (Dresden, Germany), respectively. Dr. Seo has published more than 120 articles in peer-reviewed journals, and he serves as an editorial board member of multiple journals including the Journal of Sensory Studies, Food Quality and Preference, Foods, Journal of Culinary Science and Technology, Journal of Food Science, and Korean Journal of Food and Cookery Science. He also serves as an Associate Editor and a Section Editor of the Food Research International and Current Opinion in Food Science, respectively.