

# Deep Learning Networks for View-independent Gait Events Detection

Ankhzaya Jamsrandorj<sup>b</sup>, Dawoon Jung<sup>a</sup>, Konki Sravan Kumar<sup>a</sup>, Muhammad Zeeshan Arshad<sup>a</sup>, Kyung-Ryoul Mun<sup>a</sup> and Jinwook Kim<sup>a,\*</sup>

<sup>a</sup> Center for Artificial Intelligence, Korea Institute of Science and Technology, Seoul, South Korea

<sup>b</sup> Department of Human Computer Interface & Robotics Engineering, University of Science & Technology, Daejeon, South Korea

---

## ARTICLE INFO

### Keywords:

Gait event detection  
Vision-based methods  
Deep learning  
2D convolutional neural network  
Transformer network

## ABSTRACT

Accurate gait detection is crucial in utilizing the ample health information embedded in it. As an alternative of the laborious and demanding sensor-based detection, vision-based approaches emerged. Yet, its complicate feature engineering process and heavy reliance on the lateral views serve as challenges. This study aimed to propose a view-independent vision-based gait event detection using deep learning networks that requires no pre-processing. A total of 22 participants performed seven different walking and running-related actions and the sequential video frames acquired from their actions were used as the input of the deep learning networks to produce the probability of gait events as outputs. The Transformer network and ResNet18 trained with sequential video frames achieved an F<sub>1</sub>-score of 0.956 or higher for walking straight and walking around. The detection performance on the frontal, lateral, and backside views did not differ much. The findings enhance applicability of vision-based approach and contribute to increasing its utility in health monitoring.

---

## 1. Introduction

Accurate and reliable gait assessment is crucial in utilizing the ample health information embedded in it. Ever since the introduction of inertial measurement unit (IMU) sensors and machine learning techniques, the accuracy of the sensor-based measurement improved greatly and brought about many meaningful changes in the field both academic and clinical Simonetti, Villa, Bascou, Vannozzi, Bergamini and Pillet (2019); Aqueveque, Morrison, Osorio and Pastene (2020); Gurchiek, Garabed and McGinnis (2020); Miyake, Kobayashi, Fujie and Sugano (2020); Romijnders, Warmerdam, Hansen, Welzel, Schmidt and Maetzler (2020); Bijalwan, Semwal and Mandal (2021); Sahoo, Saboo, Pratihari and Mukhopadhyay (2020); Jellish, Abbas, Ingalls, Mahant, Samanta, Ospina and Krishnamurthi (2015); Zahradka, Verma, Behboodi, Bodt, Wright and Lee (2020); Lempereur, Rousseau, Rémy-Néris, Pons, Houx, Quéllec and Brochard (2020). Still, requiring much work for a setup by a trained professional and high cooperation from a participant serves as a major challenge, hindering its widespread use in the everyday daily environment. In an effort to find a less demanding approach to detecting gait events, vision-based gait measurement has emerged and meaningful progress has been made.

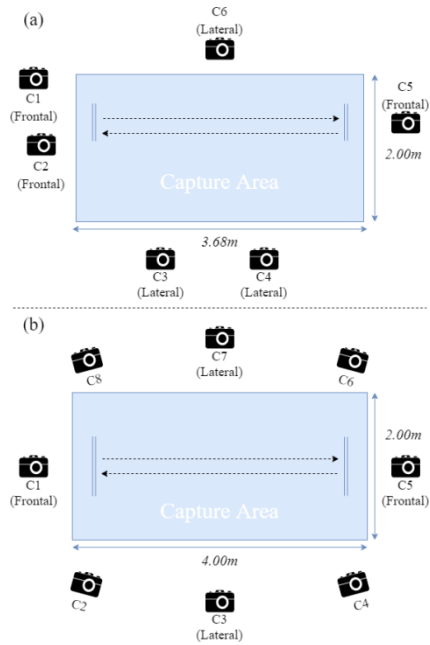
Vision-based measurement of gait solely relies on the video images to acquire data. While the appearance-based methods use silhouettes of a person to understand spatiotemporal changes between two or more subsequent frames Nieto-Hidalgo, Ferrandez, Valdivieso-Sarabia, Mora-Pascual and García-Chamizo (2015); Tang, Li, Tian, Ding and Lin (2019); Nieto-Hidalgo, Ferrández-Pastor, Valdivieso-Sarabia, Mora-Pascual and García-Chamizo (2016b); Nieto-Hidalgo, Ferrandez, Valdivieso-Sarabia, Mora-Pascual and García-Chamizo (2016a); Yang, Ugbole, Kerr, Stankovic, Stankovic, Carse, Kaliamtas and Rowe (2016); Verlekar, Vroey, Claeys, Hallez, Soares and Correia (2019), the pose-based methods use depth-sensing cameras or human pose estimation machines to acquire the skeletal structure and features such as leg length, normalized average stride, step lengths, and gait velocity Prakash, Kumar and Mittal (2016b); Prakash, Gupta, Kumar and Mittal (2016a); Arcila Cano, Ewins, Shaheen and Catalfamo Formento (2017); Rocha, Choupina, do Carmo Vilas-Boas, Fernandes and Cunha (2018); Chakraborty and Nandy (2020).

Nieto-Hidalgo *et al.* designed an appearance-based gait classification system by detecting heel-strike (HS) and toe-off (TO) events from the lateral views and achieved an F1-score of 0.83 in measuring normal gaits Nieto-Hidalgo

---

\* Corresponding author.

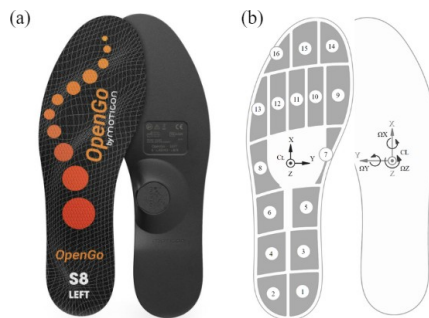
E-mail address: ankhzaya@kist.re.kr (A. Jamsrandorj), dwjung@kist.re.kr (D. Jung), konkisravan@kist.re.kr (K.S. Kumar), zeeshan@kist.re.kr (M.Z. Arshad), krmoon02@kist.re.kr (K. Mun), jwkim@imrc.kist.re.kr (J. Kim)



**Figure 1:** (a) Experimental setup with six RGB cameras for Dataset 1, (b) Experimental setup of Dataset 2 with eight RGB cameras.

et al. (2016b). Tang *et al.* proposed a new feature called consecutive silhouettes difference (CSD) maps by encoding several consecutive silhouettes to represent gait patterns and detected TO events using them with a convolutional neural network Tang et al. (2019). A markerless 2D video-based system to estimate HS and TO events has been suggested as well. As for the pose-based measurement, a fully automatic gait analysis system based on a single RGB-D camera was proposed in recognizing walking, standing, and marching activities Rocha et al. (2018) and a passive marker-based system for automated detection of HS and TO events was designed and an F1-score of 0.93 on their test set from the lateral view was achieved Prakash et al. (2016b).

Despite the inspiring achievements made, most vision-based approaches have been limited to classifying gait types such as pathological gait against normal one Verlekar, Soares and Correia (2018); Albuquerque, Machado, Verlekar, Soares and Correia (2021a); Dentamaro, Impedovo and Pirlo (2020); Sabo, Mehdizadeh, Ng, Iaboni and Taati (2020); Cao, Xue, Chen, Chen, Ma, Hu, Ma and Ma (2021); Kidziński, Yang, Hicks, Rajagopal, Delp and Schwartz (2020); Sikandar, Rabbi, Ghazali, Altwijri, Alqahtani, Almijalli, Altayyar and Ahamed (2021) or to recognizing a person by his or her gait Zhang, Wang and Li (2021); Chao, He, Zhang and Feng (2019); Singh, Jain, Arora and Singh (2018); Verlekar, Correia and Soares (2017); Liao, Cao, Garcia, Yu and Huang (2017); Shiraga, Makihara, Muramatsu, Echigo and Yagi (2016); Albuquerque, Machado, Verlekar, Correia and Soares (2021b); Albuquerque, Verlekar, Correia and Soares (2021c); Masullo, Burghardt, Damen, Perrett and Mirmehdi (2020) and little attention has been paid to gait event detection.



**Figure 2:** (a). The Moticon Science pressure sensor insole and (b). The location of sensors.

**Table 1****The demographic and anthropometric characteristics of participants of the datasets.**

	Dataset 1	Dataset 2
Number of participants (Female)	11 (3)	11 (4)
Age (years)	24.2 ± 3.8	25.8 ± 2.7
Height (cm)	170 ± 6.4	173 ± 8.5
Weight (kg)	71.7 ± 16.1	66 ± 12.5

Data are presented as mean ± SD.

Relying heavily on human silhouettes that are extracted from original RGB frames which are susceptible to the view angles and other aspects like the distance between the camera and subject or clothing conditions still serves as a major challenge of ensuring accuracy on top of feature extraction processes being laborious and time-consuming. The pose-based methods which entirely depend on depth-sensing cameras are quite costly and require an intricate set of an environment. The absence of joint information hinders the exact detection of gait as well. On top of that, vision-based approaches rely heavily on the lateral views rather than the frontal ones for the frontal ones have little contrast among the straight silhouettes and skeletons.

Hence, this study aimed to find a more reliable way of assessing gait using video images and propose a view-independent vision-based approach to detecting gait events. Attention-based deep learning networks were trained with the video images of various walking activities and predicted the foot contact (FC) and foot-off (FO) for each frame as outputs. The prediction was validated by comparing the predicted gait events with the actual gait events in frames. The findings will help increase the potential of vision-based gait assessment being applied to the everyday life of us and will contribute to the establishment of remote health monitoring (RHM) or remote diagnosis based on it.

## 2. Materials and Methods

### 2.1. Experimental protocol to obtain the two datasets

This study is based on work supported by the Institutional Review Board of the Korea Institute of Science and Technology (IRB No. 2019-029). Two experiments were conducted to obtain two separate datasets.

#### 2.1.1. Dataset 1

Eleven participants without any history or presence of neurological disorders participated in the first experiment. As shown in Fig.1(a), the data was collected from an indoor environment under 6 different viewing angles: 3 frontal views and 3 lateral views. The capture area had a size of 3.68 meters in length and 2.0 meters in width. The data from a participant was simultaneously captured by timely-synchronized 6 RGB cameras with a resolution of Full HD and a frame rate of 60 Hz. There were three different walking actions in this dataset: *Walking Straight (WS)*, *Walking Around (WA)*, and *Walking on Treadmill (WoT)*. For WS, each participant was requested to walk straight at their self-selected pace and turn back at a specific point and repeat this for a minute. For WA, participants were asked to walk around freely at their self-selected pace in the capture area for a minute as well. As for WoT, participants were asked to walk on a treadmill for 30 seconds with its speed fixed at 3.5m/s. The participants were requested to carry out each action twice. A total of 1.8M frames were acquired in the dataset. The dataset details are summarized in Table 2.

#### 2.1.2. Dataset 2

Another eleven participants were involved in the second experiment. The data was collected from an indoor environment, as shown in Fig.1(b), and the capture area had a size of 4.0 and 2.0 meters. Timely-synchronized 8 RGB cameras simultaneously captured the data with a resolution of 1280×720 and a frame rate of 60 Hz. The participants were instructed to perform seven walking and running-related actions at three different speeds: *WS*, *WA*, *Running Straight (RS)*, *Running Around (RA)*, *Walking on Toes (WT)*, *Running on Toes (RT)* and *Walking in Place (WIP)* at preferred, slow, and fast speeds. The participants were asked to perform all actions 15% to 25% slower and faster than their preferred speed for slow and fast speeds. For WS and RS actions, each participant was requested to walk and run straight and turn back at a specific point at preferred, slow, and fast speeds for 35 seconds. As for WA and RA actions, participants were asked to walk and run around freely in the capture area at three different speeds for 35 seconds as well.

**Table 2**  
**The summary of the Datasets for Gait Event Detection**

	No of subjects	Action name	Time	Speeds	Frame rate	No of Views	Total frames
Dataset 1	11 (male 8 and female 3)	Walking Straight	60 x 2	n	60fps	6	475200
		Walking Around					475200
		Walking on Treadmill	30 x 2	3.5m/s			237600
<b>Dataset 1 size</b>							<b>1188000</b>
Dataset 2	11 (male 7 and female 4)	Walking Straight			60fps	8	1108800
		Walking Around					1108800
		Running Straight					1108800
		Running Around	35 x 2	s/n/f			1108800
		Walk on Toes					1108800
		Run on Toes					1108800
		Walking in Place					1108800
<b>Dataset 2 size</b>							<b>7761600</b>

Walking speeds were self-selected: preferred (n), slow (s), and fast (f).

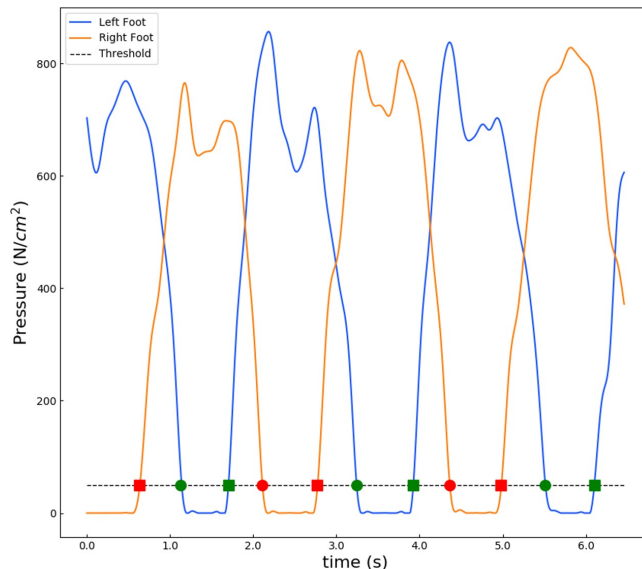
For WT and RT actions, the participants were instructed to walk and run only on their toes for 35 seconds without their heels contacting on the ground. The participants were requested to carry out each action twice. The dataset acquired was 6.5 times bigger than the first dataset. Table 2 summarized the dataset details.

All 22 participants provided a written informed consent form prior to the experiments. All participants neither reported inconvenience while walking nor had gait disturbance identifiable by naked eyes. The demographic and anthropometric characteristics of the participants are summarized in Table 1. Each video was equipped with human bounding boxes on every frame and annotations containing the frame numbers in four gait events: right FC and FO and left FC and FO.

## 2.2. Target construction for gait events detection

### 2.2.1. Gait temporal parameters extraction

In the medical field, FC event is defined as the moment that foot touches the ground while FO event is defined as the moment that foot leaves the ground (Fig.4(a)). Following above definition, all the video frames acquired from the dataset 1 were manually labeled by an expert annotating FC and FO events of both legs. The labels then were cross-checked by another expert.



**Figure 3:** Gait events detection from the total pressure data measured at the left and right foot. The threshold-crossing points respectively determined the foot contact and foot off times. (FC–squares, FO–circles)

**Table 3**
**The condition checks of the developed gait event detection algorithm for labeling.**

	Label	Condition Checks & Obtain frame numbers
FC	Time-stamps	$[(TP_{i+1} < TP_i < TP_{i+1}) \wedge \min( TP_i - Th ,  TP_{i+1} - Th )]$
	Frame Numbers	$[TS \times FPS]$
FO	Time-stamps	$[(TP_{i+1} > TP_i > TP_{i+1}) \wedge \min( TP_i - Th ,  TP_{i+1} - Th )]$
	Frame Numbers	$[TS \times FPS]$

$TP_i$  is the  $i^{\text{th}}$  value of total pressure data and  $Th$  is the pressure threshold.

$TS$  is the values of the time-stamp defined with previous conditions, and  $FPS$  is a frame rate.

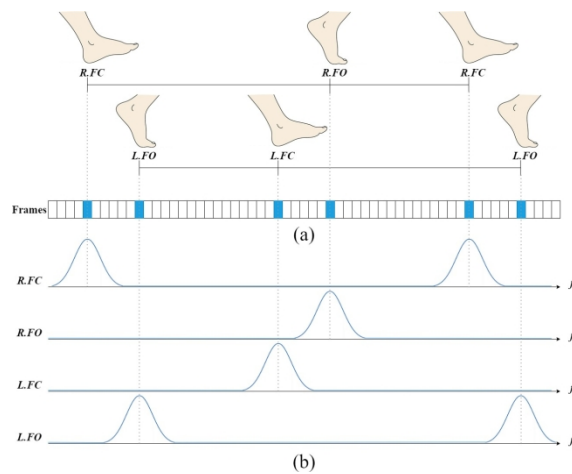
For dataset 2, the Moticon SCIENCE pressure sensor insoles were used to reduce the labeling time and costs Morin, Muller, Pontonnier and Dumont (2021); Antwi-Afari, Yu, Li, Darko, Seo and Wong (2018); Chatzaki, Skaramagkas, Tachos, Christodoulakis, Maniadi, Kefalopoulou, Fotiadis and Tsiknakis (2021). The Moticon SCIENCE sensor insole (Fig.2(a)) is a professional, wireless measurement tool that has a sufficient number of pressure sensors (Fig.2(b)) and a 6-axis Inertial Measurement Unit (IMU) sensor. These pressure insoles were inserted inside the shoes of each participant. The annotation of gait events was made using the gait event detection algorithm developed for this purpose. The gait event detection algorithm was based on the total pressure data from smart sensor insoles for both legs and the pressure threshold parameter which was computed as follows where  $W$  is the bodyweight of a participant and  $F_p$  is a pressure factor of 0.98:

$$\text{Threshold} = W \times F_p \quad (1)$$

As Fig.3 shows, the time-stamps in which gait events occur were extracted by analyzing total force data for both legs. The closest point respectively determined the FC and FO times with the pressure threshold, and these time-stamps were multiplied by a frame rate of 60Hz to obtain the frame numbers that contain the FC and FO events. Table 3 shows the condition checks and the respective labels of the developed gait event detection algorithm for this labeling. The results of the algorithm were cross-validated using the video data.

### 2.2.2. Target construction

To avoid the training becoming unstable by micro labelling of each gait event which makes the labels too sparse for frame numbers, the hard labels were smoothed into soft labels. We smoothed our raw and frame-specific labels to produce probability distributions that fitted around the frames in which gait events occurred since the neighboring frames of each FC and FO event having similar pixel contents could confuse the networks and burden the learning. The target was four one-dimensional Gaussian distribution curves that were associated with the four gait events, respectively (Fig.4(b)).



**Figure 4:** (a) Definitions used to annotate gait events on video frames and (b) Target construction for gait events detection. Note: FC can be heel-strike, midfoot-strike, and toe-strike, and FO can be toe-off, midfoot-off, and heel-off.

### 2.3. Network architectures

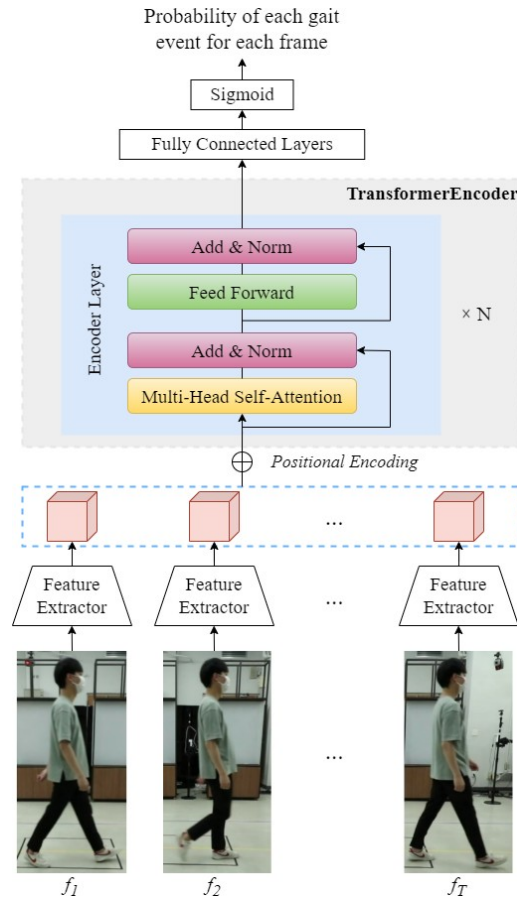
Two different deep learning networks were developed for gait events detection: Attention-based and 2D convolutional neural network (CNN) networks.

#### 2.3.1. Attention-based network

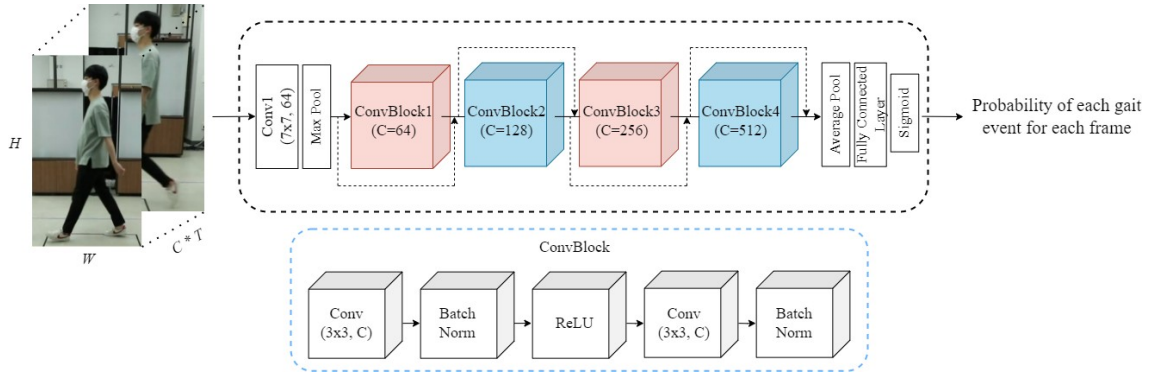
The architecture of the attention-based network was modular and composed of three consecutive parts: a 2D spatial feature extraction model, a temporal attention-based encoder, and an MLP-based detector. Fig.5 demonstrates the architecture of the attention-based network. We adopted the ResNet18 pre-trained on ImageNet as the backbone spatial feature extractor for its relatively small size and proven effectiveness. A total of 27 subsequent frames were used as the input. We used a Transformer encoder architecture that applies the attention mechanisms for temporal feature extraction. The sequence of feature vectors injected positional information and then fed it to the Transformer encoder, which contained two encoder layers. Each encoder layer had four multi-head self-attention and a feed-forward network. A final module, the MLP-based detector consisted of two fully connected layers with a ReLU activation function and Dropout between them.

#### 2.3.2. 2D convolutional neural network

The architecture of the 2D CNN network used in this study is depicted in Fig.6. A stack of 9 subsequent frames was used as the input of the network for obtaining the spatiotemporal and motion representations, and the output was the probability of each gait event for each frame. The experiment was carried out with ResNet18. For this network, the last FC layer was removed and a new FC layer was added to adapt to our output.



**Figure 5:** The architecture of the attention-based network. The Transformer encoder, which consisted of two Encoder layers, was used. Each encoder layer had four multi-head self-attention sublayers and a feed-forward network.



**Figure 6:** The architecture of the two-dimensional convolutional neural network. The ResNet18 pre-trained on ImageNet was used. Input is a stack of  $T$  frames, and output is the probability of each gait event for each frame.

## 2.4. Implementation details

### 2.4.1. Data processing

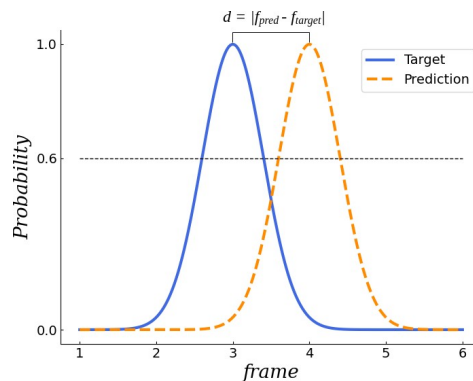
The input tensors were constructed from  $T$  subsequent frames with the ground truth human bounding boxes and each event probability target for all  $T$  frames. The raw frames were pre-processed by cropping the person in the frame based on the ground truth human bounding box, then pad left and right to get a square image and resized it to a size of  $448 \times 448$ . The dataset was separated by participants into three parts: the data from 11 participants for the training set (1036.8k frames), data from 6 participants for the validation set (561.6k frames), and the data from the remaining 5 participants for the test set (460.8k frames). We used data from WS and WA actions for the training, validation, and testing in the experiments, and the rest of the actions were used for an additional test.

### 2.4.2. Training scheme

All experiments were carried out with the PyTorch framework. The weights of the networks were initialized by pre-training on the ImageNet for gait events detection. During all training experiments, the Mean Squared Error (MSE) and the Adam optimizer were used with an initial learning rate of 0.001 and a batch size of 16. The learning rate was reduced two times after every three epochs. The horizontal flip was applied to input frames with a probability of 0.5.

## 2.5. Evaluation metrics

The target for gait events detection was designed as probability distributions for each gait event, and the detection was assessed with two metrics. A Smooth Percentage Correct Events (SPCE) which calculates the difference between target and prediction probabilities for every frame was adopted Voeikov, Falaleev and Baikulov (2020).



**Figure 7:** Graphical demonstration of frame-error that used to calculate the precision of the prediction. The frame difference between the target frame of gait events and the prediction frame is noted as  $d$ . To discretize the probability distribution into precise frame numbers that include gait events, the peak values were picked at a threshold of 0.6.

**Table 4**  
The performances of the networks on the test set of Dataset 1.

	Model	Frontal		Lateral	
		SPCE	F1	SPCE	F1
WS	ResNet18	0.906 ± 0.005	0.901 ± 0.008	0.908 ± 0.004	0.884 ± 0.005
	Transformer	0.938 ± 0.004	0.912 ± 0.020	0.942 ± 0.002	0.921 ± 0.002
WA	ResNet18	0.926 ± 0.009	0.956 ± 0.007	0.918 ± 0.005	0.959 ± 0.004
	Transformer	0.943 ± 0.007	0.961 ± 0.010	0.947 ± 0.005	0.964 ± 0.002
WoT	ResNet18	0.921 ± 0.010	0.961 ± 0.039	0.828 ± 0.102	0.911 ± 0.041
	Transformer	0.937 ± 0.009	0.993 ± 0.003	0.923 ± 0.009	0.979 ± 0.017

Data are presented as mean ± SD.

An event was considered correct if this probability difference was less than a threshold of 0.25. The second metric was the F<sub>1</sub>-score (**F1**) which evaluates the precision of the gait event detection which was calculated as follows:

$$F_1 - score = 2 \times \frac{P \times R}{P + R} \quad (2)$$

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$R = \frac{TP}{TP + FN} \quad (4)$$

where  $P \in (0, 1)$  and  $R \in (0, 1)$  are precision and recall, respectively. TPs are true positives, FPs are false positives, and FNs are false negatives. As shown in Fig.7, for each probability distribution, the peak values were first picked at the threshold of 0.6 to precise the frame numbers containing gait events, then the frame difference between the target frame number and the prediction frame number was calculated. Each event was treated as a TP if the frame difference was less than a threshold of 4 frames. All other gait event predictions were considered as FPs, and all targets that were not detected were treated as FNs. These TPs, FPs, and FNs were used to calculate the precision showing how many detected gait events matched with the actual gait events and the recall showing how many actual gait events were detected.

### 3. Results

The distributions of the probabilities for each gait event were illustrated in Fig.8, which shows that both networks properly detected the frames with FC and FO events.

Table 4 summarizes the performance of the networks on WS and WA actions in Dataset 1. The performances on frontal and lateral views were reported separately. As shown in Table 4, the Transformer network achieved the best performance in detecting the gait events for both WS and WA actions. For WS action, an F<sub>1</sub>-score of 0.912 (precision 0.948 and recall 0.877) was achieved from the frontal view and that of 0.921 (precision 0.952 and recall 0.892) was achieved from the lateral view. As for WA action, an F<sub>1</sub>-score of 0.961 (precision 0.955 and recall 0.970) was achieved from the frontal view and that of 0.964 (precision 0.957 and recall 0.974) was achieved from the lateral views. The highest values of SPCE for WS action were 0.938 and 0.942 from the frontal and lateral views while those for WA action were 0.943 and 0.947 from the frontal and lateral views.

Table 5 shows the performance of the networks on the test set of Dataset 2. The Transformer network for WS action at the preferred-, slow-, and fast-paced walking achieved the highest F<sub>1</sub>-scores of 0.904 (precision 0.966 and recall 0.846), 0.898 (precision 0.959 and recall 0.833), and 0.960 (precision 0.975 and recall 0.944), respectively. For WA action, the highest F<sub>1</sub>-scores of 0.982 (precision 0.973 and recall 0.986), 0.987 (precision 0.996 and recall 0.978), and 0.988 (precision 0.983 and recall 0.996) were achieved at the preferred-, slow-, and fast-paced walking. The Transformer network on the test set of Dataset 2 achieved the highest SPCE values for both WS and WA actions. As Table 5 shows, the highest SPCE values for WS action were 0.934, 0.924, and 0.947 at the preferred-, slow-, and fast-paced walking, whereas, those for WA action were 0.958, 0.973, and 0.953.

The performance of the networks using the WT and RA actions from Dataset 2 and WoT action from Dataset 1 was examined. Although these actions were not used for training, the Transformer network achieved an  $F_1$ -score of 0.993 (precision 0.993 and recall 0.991) for WoT from the frontal view with the highest SPCE of 0.937. As for the lateral view of the same action, an  $F_1$ -score of 0.979 (precision 0.990 and recall 0.973) was achieved with the highest SPCE of 0.923 (Table 4). As for the WT action, the Transformer network achieved an  $F_1$ -score of 0.952 (precision 0.956 and recall 0.948) in detecting gait events at preferred-paced and 0.998 (precision 0.997 and recall 0.997) at fast-paced walking (Table 5). For WT action at slow speed, the ResNet18 network achieved the highest detection performance showing an  $F_1$ -score of 0.942 (precision 0.946 and recall 0.936). For RA actions, the Transformer network achieved an  $F_1$ -scores of 0.977 (precision 0.990 and recall 0.960) at the preferred speed, while that of 0.986 (precision 0.999 and 0.976) and 0.964 (precision 0.998 and 0.937) were achieved at slow and fast speed. The highest SPCEs for this action at the preferred-, slow-, and fast-speed were 0.958, 0.973, and 0.953, respectively.

Table 6 and Table 7 summarises FO and FC events detection results of the two networks on the test set of Dataset 1 and Dataset 2, respectively. For both ResNet18 and Transformer networks, the detection performance of the FO event was higher than that of the FC event. The deep learning networks achieved the average  $F_1$ -scores of 0.905 and 0.934 for the left and right FO event detection for WS action while those for the right and left FC event detection were 0.900 and 0.876. For WA action, the average  $F_1$ -scores were 0.975 and 0.967 for the left and right FO event detection, and those for the left and right FC event detection were 0.947 and 0.923.

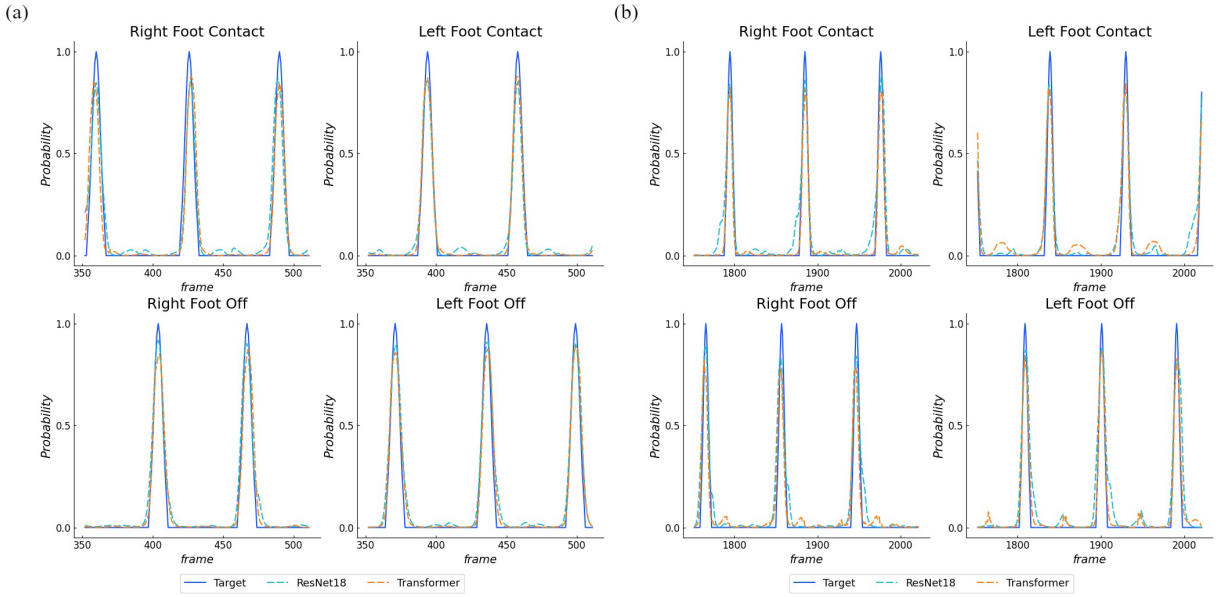
As illustrated in Table 7, for WS action of Dataset 2, the average  $F_1$ -scores detecting the left and right FO events were 0.930 and 0.923. The average detection performance of the networks was  $F_1$ -scores of 0.885 and 0.898 on the left and right FC events. For WA action, the average detection performance was  $F_1$ -scores of 0.989 and 0.991 for the left and right FO event detection, and those for the left and right FC event detection were 0.960 and 0.978. The Attention-based network achieved the best performance for most actions, while a simple 2D CNN network trained with a stack of nine frames likewise achieved satisfactory results.

**Table 5**  
The performances of the networks on the test set of Dataset 2.

	Models	Walking Speed	SPCE	F1
WS	ResNet18	N	0.901 ± 0.011	0.889 ± 0.017
		S	0.903 ± 0.012	0.868 ± 0.030
		F	0.936 ± 0.005	0.939 ± 0.015
	Transformer	N	0.934 ± 0.006	0.904 ± 0.031
		S	0.924 ± 0.010	0.898 ± 0.030
		F	0.947 ± 0.008	0.960 ± 0.015
WA	ResNet18	N	0.952 ± 0.006	0.973 ± 0.011
		S	0.954 ± 0.008	0.979 ± 0.010
		F	0.942 ± 0.007	0.971 ± 0.012
	Transformer	N	0.958 ± 0.014	0.982 ± 0.004
		S	0.973 ± 0.005	0.987 ± 0.011
		F	0.953 ± 0.007	0.988 ± 0.004
WT	ResNet18	N	0.881 ± 0.011	0.940 ± 0.028
		S	0.916 ± 0.009	0.942 ± 0.020
		F	0.959 ± 0.008	0.985 ± 0.012
	Transformer	N	0.904 ± 0.014	0.952 ± 0.022
		S	0.933 ± 0.010	0.933 ± 0.036
		F	0.966 ± 0.004	0.998 ± 0.003
RA	ResNet18	N	0.918 ± 0.007	0.940 ± 0.020
		S	0.951 ± 0.009	0.971 ± 0.016
		F	0.905 ± 0.010	0.915 ± 0.012
	Transformer	N	0.939 ± 0.009	0.977 ± 0.013
		S	0.973 ± 0.009	0.986 ± 0.007
		F	0.922 ± 0.012	0.964 ± 0.020

Walking speeds: preferred (N), slow (S), and fast (F).

Data are presented as mean ± SD.



**Figure 8:** (a). The probability distributions of detected and actual gait events for WS action and (b). The probability distributions of detected and actual gait events for WA action.

## 4. Discussion

This study pioneered a way of improving the reliability of vision-based gait detection and proposed a view-independent approach. Training attention-based deep learning networks with sequential video frames achieved outstanding detection performance regardless of the camera view angles. On top of improving reliability, the proposed approach requires neither a sophisticated experimental setup nor time-consuming pre-processing, advancing its applicability greatly from the previous vision-based gait detection approaches. When applied in practice, the proposed approach can provide a great convenience in terms of its use. With the videos of their walk, patients can communicate with their physicians whether they are in need of making a trip to a practice or hospital in advance which can prevent unnecessary trips. Since filming does not require any professional knowledge or intricate engineering, it can serve as a simple and burden-free monitoring tool for anyone. Moreover, it can serve as an alternative to a self-reported survey when investigating the health conditions of the elderly who are having a hard time addressing their issues properly for their declining cognitive ability. The potential of the proposed approach in the field of future RHM or remote diagnosis based on it can be found enormous.

The study made a step forward from the previous studies that explored vision-based gait detection by making it less view-dependent and improving its accuracy. Previously, the studies on vision-based gait detection mainly used human silhouettes and human pose information extracted from original RGB frames. The challenge was that those silhouettes altered easily depending on the view angles and distance between the camera and human. Clothing conditions served as another obstacle as well. Nieto-Hidalgo *et al.* tried to break free from the heavy reliance on the lateral images

**Table 6**  
Comparison of detection results on FO and FC events for the actions of Dataset 1.

	Model	F1			
		LFC	LFO	RFC	RFO
WS	ResNet18	0.891	0.884	0.871	0.925
	Transformer	0.909	0.927	0.880	0.944
WA	ResNet18	0.941	0.974	0.921	0.967
	Transformer	0.953	0.977	0.926	0.968

**Table 7**  
**Comparison of detection results on FO and FC events for the actions of Dataset 2.**

	Model	Walking speed	F1			
			LFC	LFO	RFC	RFO
WS	ResNet18	N	0.842	0.906	0.879	0.928
		S	0.865	0.901	0.858	0.846
		F	0.888	0.939	0.950	0.975
	Transformer	N	0.891	0.923	0.881	0.920
		S	0.906	0.922	0.858	0.904
		F	0.919	0.989	0.961	0.967
WA	ResNet18	N	0.938	0.982	0.979	0.986
		S	0.963	0.996	0.968	0.993
		F	0.932	0.983	0.979	0.988
	Transformer	N	0.969	0.986	0.981	0.991
		S	0.986	0.998	0.969	0.991
		F	0.974	0.990	0.991	1.000
WT	ResNet18	N	0.933	0.959	0.898	0.971
		S	0.896	0.986	0.917	0.965
		F	0.964	0.993	0.993	0.994
	Transformer	N	0.916	0.990	0.906	0.996
		S	0.878	0.991	0.902	0.958
		F	0.988	0.996	1.000	0.998
RA	ResNet18	N	0.859	0.969	0.964	0.969
		S	0.904	0.982	0.990	0.994
		F	0.786	0.969	0.927	0.979
	Transformer	N	0.959	0.993	0.976	0.998
		S	0.975	0.998	0.974	1.000
		F	0.952	0.994	0.930	0.995

Walking speeds: preferred (N), slow (S), and

fast (F).

and suggested a gait analysis system based on the frontal images. Yet, the highest accuracy achieved was 89.1% which challenged its application in real practice Nieto-Hidalgo et al. (2016a). The accuracy achieved by Xu *et al.* who attempted to utilize the frontal images for vision-based gait detection was 94.3% for HS events and 42.7% for TO events Xu, McGorry, Chou, Lin and Chang (2015). The accuracy achieved by the Transformer network in this study far exceeds those obtained by these studies, increasing the reliability and potential of the vision-based gait detection. To our knowledge, this is the first study that invited multiple view angles in one. Lacking information, previous studies using only one view angle called for an additional procedure of acquiring additional information but this study managed to obtain a high accuracy using RGB frames exclusively while considering three different angles at the same time. On top of that, it should be worth noting that this is the first study that investigated the use of backside views for vision-based gait detection. Despite years of research, little attention has been paid to the use of backside view for gait detection. The results of this study show the detection performance for the lateral, frontal, and backside views did not differ significantly, suggesting that the backside view can be comparable to the frontal and lateral ones.

The outstanding detection performance that the Transformer network achieved in this study might have derived from its ability to utilize the sequential data. Since gait events are sequential, the networks were required to learn the dependencies between the two sequential events. The simple CNN, ResNet18 here, can learn these dependencies using different kernels, however, the computational cost would be enormous when kernel size increases to capture dependencies of long-range sequences. Unlike CNN, the transformer network can capture dependencies from long-range input sequences using multi-head attention mechanisms and positional encodings, avoiding any possible computational cost problems. Using 27 frames as inputs which were far more than 9 of the ResNet18 network, the Transformer network might have been able to outperform the other in its detection. Having more steps of feature extraction might have contributed to its better detection as well.

The F<sub>1</sub>-scores achieved with WS and WA actions varied, showing better detection performance with WA action compared to that with WS action. A possible explanation behind this can be that, unlike WA action in which the participants did not stop walking, WS action had *walking straight* and *turning-back* stages and the detection of these two different stages varied. While gait events were well detected in the walking straight stage, the detection dropped significantly in the turning back stage (Table 8).

**Table 8**  
**Comparison of the performances of the networks at different motions.**

Model	F1	
	Straight Walking Stage	Turning Back Stage
ResNet18	0.987 ± 0.004	0.634 ± 0.059
Transformer	0.992 ± 0.003	0.708 ± 0.046

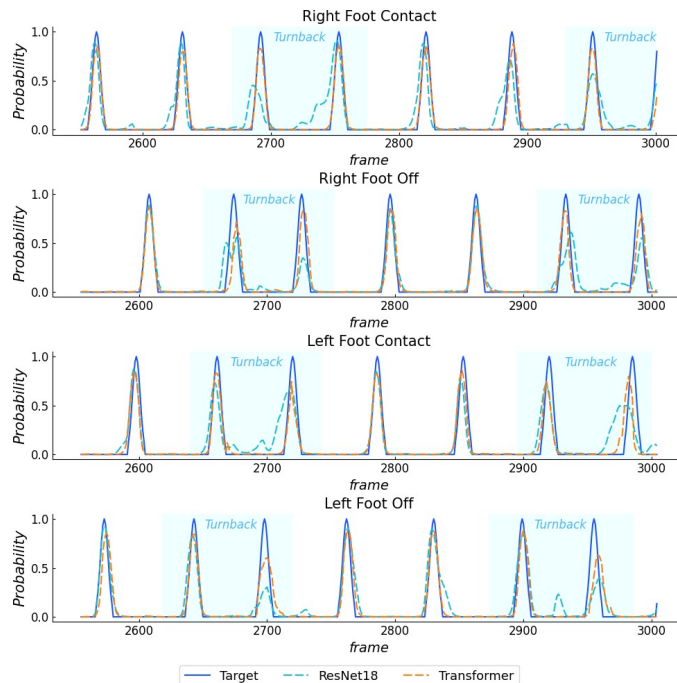
Data are presented as mean ± SD.

The detection performance of FC events was lower than that of FO events in the turning back stage as well. The act of turning back which involves some self-occlusions between the left and right limbs may have affected the lowering of the detection performance. Nevertheless, the Transformer network made a quite reliable detection even in the turning back stage (Fig.9).

The networks in this study showed a fairly good detection of RA and WT actions even though they had not been introduced to the data for these actions. Detecting gait events in RA and WT actions has been a major challenge for the conventional detection approaches based on HS and TO since WT action does not exhibit any HS for it starts from toe-strike and ends with toe-off and RA action has fewer HS as it involves a midfoot-strike with more toe-strike as the running speed increases. Using FC and FO alternatively may have contributed to better detection of these actions. It is quite remarkable that the networks detected gait events in RA action accurately with its limited training with walking-related actions only.

When the detection performance of using the hard labels was compared with that of using the soft labels, the detection performance was better with the soft labels (Table 9). Training the networks with the hard labels landed in a relatively poor detection due to their learning being challenged by too many frames marked as 0. Softening the targets to smooth probability distribution may have reduced the burden of learning which in turn resulted in better detection performance.

The study found a mild drop in the detection performance at slow-paced walking compared with that at preferred- and fast-paced walking. As shown in Table 10, the detection performance did not differ greatly by the number of thresholds used but as for the slow-paced walking, a great variation by thresholds numbers was observed. When participants walked slowly, they not only walked slowly but also waited for some time before making the following action.



**Figure 9:** Comparison of detection results at different motions.

**Table 9**  
**Comparison of the performances of the networks trained with the Hard- and Soft-Labels.**

	Model	F1	
		Hard-Label	Soft-Label
WS	ResNet18	0.590 ± 0.098	0.893 ± 0.011
	Transformer	0.708 ± 0.018	0.917 ± 0.013
WA	ResNet18	0.683 ± 0.048	0.957 ± 0.005
	Transformer	0.739 ± 0.038	0.963 ± 0.006

Data are presented as mean ± SD.

Therefore more frames between the two sequential events were acquired which might have burdened the learning by requiring more time for the networks to recognize the features of the coming events. The detection performance for slow-paced walking improved when a more comprehensive threshold was used.

Despite the remarkable breakthrough it has made, the study bears several limitations. The first is that the study population was exclusively young and healthy so the networks trained and tested were void of the gaits of the elderly or ones with any health issues. Further studies including various age groups and health conditions such as Parkinson's disease, osteoarthritis, stroke, cerebral palsy, multiple sclerosis, and muscular dystrophy should follow to validate the proposed approach. Secondly, the experiments for this study were conducted under a well-controlled situation in a limited space. Using a well-organized and rather rectangular shape of a space, the study is in no position to verify whether the same results can be achieved under long corridor experimental setups where the scale of people varies greatly depending on the distance from the camera. Future validation of this study at various experimental setups should be carried out. The third is that, although the heights of the participants were different, the height of the cameras was fixed and no other adjustment according to the heights of a participant was made. For this group of participants, however, the heights of the participants did not deviate significantly from the average range. For the participants whose heights vary greatly, it may be necessary to adjust the height of the camera depending on their heights.

## 5. Conclusion

This study aimed to find a more reliable vision-based gait event detection and found an approach that is view-independent and accurate. Training Attention-based and 2D CNN networks with sequential video frames, the Transformer network detected gait events with the highest  $F_1$ -score of 0.998. The study has also shown that the use of the frontal view is comparable to using the lateral view. The applicability of the proposed approach can be enormous given that it does not require any special camera setup or complicate feature engineering process. Future studies that validate the proposed approach with the gaits from various health conditions such as Parkinson's disease, frailty, and cognitive impairment should follow. The findings can serve as one of the first stepping stones towards future studies for gait-based health monitoring both at home and in a clinical setting.

**Table 10**  
 **$F_1$ -scores of the networks at different thresholds.**

	Model	Walking Speed	F1		
			Thresh=2 (33 ms)	Thresh=4 (66 ms)	Thresh=6 (100 ms)
WS	ResNet18	N	0.749 ± 0.032	0.889 ± 0.017	0.906 ± 0.020
		S	0.626 ± 0.058	0.868 ± 0.030	0.913 ± 0.027
		F	0.874 ± 0.033	0.939 ± 0.015	0.943 ± 0.015
	Transformer	N	0.751 ± 0.022	0.904 ± 0.031	0.918 ± 0.030
		S	0.731 ± 0.095	0.898 ± 0.030	0.919 ± 0.027
		F	0.852 ± 0.032	0.960 ± 0.015	0.974 ± 0.016
WA	ResNet18	N	0.929 ± 0.032	0.973 ± 0.011	0.974 ± 0.016
		S	0.903 ± 0.040	0.979 ± 0.010	0.979 ± 0.015
		F	0.918 ± 0.027	0.971 ± 0.012	0.973 ± 0.011
	Transformer	N	0.927 ± 0.018	0.982 ± 0.004	0.984 ± 0.006
		S	0.907 ± 0.020	0.987 ± 0.011	0.989 ± 0.009
		F	0.929 ± 0.039	0.988 ± 0.004	0.989 ± 0.004

Walking speeds: preferred (N), slow (S), and fast (F).  
 Data are presented as mean ± SD.

## Acknowledgment

This research was supported in part by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project No. 1711139131) and in part by the High-Tech Based National Athletic Performance Improvement (Winter) Project from the Korea Sports Promotion Foundation (Project No. B0081219000176).

## References

- Albuquerque, P., Machado, J., Verlekar, T., Soares, L., Correia, P., 2021a. Remote Pathological Gait Classification System.
- Albuquerque, P., Machado, J.P., Verlekar, T.T., Correia, P.L., Soares, L.D., 2021b. Remote gait type classification system using markerless 2d video. *Diagnostics* 11.
- Albuquerque, P., Verlekar, T.T., Correia, P.L., Soares, L.D., 2021c. A spatiotemporal deep learning approach for automatic pathological gait classification. *Sensors* 21. URL: <https://www.mdpi.com/1424-8220/21/18/6202>, doi:10.3390/s21186202.
- Antwi-Afari, M., Yu, Y., Li, H., Darko, A., Seo, J., Wong, A., 2018. Automated Detection and Classification of Construction Workers' Awkward Working Postures using Wearable Insole Pressure Sensors.
- Aqueveque, P., Morrison, E.G., Osorio, R., Pastene, F., 2020. Gait segmentation method using a plantar pressure measurement system with custom made capacitive sensors. *Sensors* 20, 656. doi:10.3390/s20030656.
- Arcila Cano, A., Ewins, D., Shaheen, A., Catalfamo Formento, P., 2017. Evaluation of methods based on conventional videography for detection of gait events, in: Torres, I., Bustamante, J., Sierra, D.A. (Eds.), VII Latin American Congress on Biomedical Engineering CLAIB 2016, Bucaramanga, Santander, Colombia, October 26th -28th, 2016, Springer Singapore, Singapore. pp. 181–184.
- Bijalwan, V., Semwal, V.B., Mandal, T.K., 2021. Fusion of multi-sensor-based biomechanical gait analysis using vision and wearable sensor. *IEEE Sensors Journal* 21, 14213–14220. doi:10.1109/JSEN.2021.3066473.
- Cao, X., Xue, Y., Chen, J., Chen, X., Ma, Y., Hu, C., Ma, H., Ma, H., 2021. Video based shuffling step detection for parkinsonian patients using 3d convolution. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* PP, 1. doi:10.1109/TNSRE.2021.3062416.
- Chakraborty, S., Nandy, A., 2020. An unsupervised approach for gait phase detection, in: 2020 4th International Conference on Computational Intelligence and Networks (CINE), pp. 1–5. doi:10.1109/CINE48825.2020.234396.
- Chao, H., He, Y., Zhang, J., Feng, J., 2019. Gaitset: Regarding gait as a set for cross-view gait recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 8126–8133. doi:10.1609/aaai.v33i01.33018126.
- Chatzaki, C., Skaramagkas, V., Tachos, N., Christodoulakis, G., Maniadi, E., Kefalopoulou, Z., Fotiadis, D.I., Tsiknakis, M., 2021. The smart-insole dataset: Gait analysis using wearable sensors with a focus on elderly and parkinson's patients. *Sensors* 21. URL: <https://www.mdpi.com/1424-8220/21/8/2821>, doi:10.3390/s21082821.
- Dentamaro, V., Impedovo, D., Pirlo, G., 2020. Gait analysis for early neurodegenerative diseases classification through the kinematic theory of rapid human movements. *IEEE Access* 8, 193966–193980. doi:10.1109/ACCESS.2020.3032202.
- Gurchiek, R., Garabed, C., McGinnis, R., 2020. Gait event detection using a thigh-worn accelerometer. *Gait Posture* 80. doi:10.1016/j.gaitpost.2020.06.004.
- Jellish, J., Abbas, J., Ingalls, T., Mahant, P., Samanta, J., Ospina, M., Krishnamurthi, N., 2015. A system for real-time feedback to improve gait and posture in parkinson's disease. *IEEE Journal of Biomedical and Health Informatics* 19, 1. doi:10.1109/JBHI.2015.2472560.
- Kidziński, , Yang, B., Hicks, J., Rajagopal, A., Delp, S., Schwartz, M., 2020. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature Communications* 11, 4054. doi:10.1038/s41467-020-17807-z.
- Lempereur, M., Rousseau, F., Rémy-Néris, O., Pons, C., Houx, L., Quellec, G., Brochard, S., 2020. A new deep learning-based method for the detection of gait events in children with gait disorders: Proof-of-concept and concurrent validity. *Journal of Biomechanics* 98, 109490. URL: <https://www.sciencedirect.com/science/article/pii/S0021929019307407>, doi:<https://doi.org/10.1016/j.jbiomech.2019.109490>.
- Liao, R., Cao, C., Garcia, E., Yu, S., Huang, Y., 2017. Pose-Based Temporal-Spatial Network (PTSN) for Gait Recognition with Carrying and Clothing Variations. doi:10.1007/978-3-319-69923-3\_51.
- Masullo, A., Burghardt, T., Damen, D., Perrett, T., Mirmehdi, M., 2020. Person re-id by fusion of video silhouettes and wearable signals for home monitoring applications. *Sensors* 20. URL: <https://www.mdpi.com/1424-8220/20/9/2576>, doi:10.3390/s20092576.
- Miyake, T., Kobayashi, Y., Fujie, M., Sugano, S., 2020. Gait event detection based on inter-joint coordination using only angular information. *Advanced Robotics* 34, 1–11. doi:10.1080/01691864.2020.1803126.
- Morin, P., Muller, A., Pontonnier, C., Dumont, G., 2021. Foot contact detection through pressure insoles for the estimation of external forces and moments: application to running and walking. *Computer Methods in Biomechanics and Biomedical Engineering* .
- Nieto-Hidalgo, M., Ferrandez, J., Valdivieso-Sarabia, R., Mora-Pascual, J., García-Chamizo, J., 2015. Vision Based Extraction of Dynamic Gait Features Focused on Feet Movement Using RGB Camera. doi:10.1007/978-3-319-26508-7\_16.
- Nieto-Hidalgo, M., Ferrandez, J., Valdivieso-Sarabia, R., Mora-Pascual, J., García-Chamizo, J., 2016a. Vision Based Gait Analysis for Frontal View Gait Sequences Using RGB Camera. doi:10.1007/978-3-319-48746-5\_3.
- Nieto-Hidalgo, M., Ferrández-Pastor, F.J., Valdivieso-Sarabia, R.J., Mora-Pascual, J., García-Chamizo, J.M., 2016b. A vision based proposal for classification of normal and abnormal gait using rgb camera. *Journal of Biomedical Informatics* 63, 82–89. URL: <https://www.sciencedirect.com/science/article/pii/S1532046416300806>, doi:<https://doi.org/10.1016/j.jbi.2016.08.003>.

- Prakash, C., Gupta, K., Kumar, R., Mittal, N., 2016a. Fuzzy Logic-Based Gait Phase Detection Using Passive Markers. pp. 561–572. doi:10.1007/978-981-10-0448-3\_46.
- Prakash, C., Kumar, R., Mittal, N., 2016b. Automated detection of human gait events from conventional videography. doi:10.1109/ETCT.2016.7882987.
- Rocha, A.P., Choupina, H.M.P., do Carmo Vilas-Boas, M., Fernandes, J.M., Cunha, J.P.S., 2018. System for automatic gait analysis based on a single rgb-d camera. PLOS ONE 13, e0201728–. URL: <https://doi.org/10.1371/journal.pone.0201728>.
- Romijnders, R., Warmerdam, E., Hansen, C., Welzel, J., Schmidt, G., Maetzler, W., 2020. Validation of IMU-Based Gait Event Detection During Curved Walking and Turning in Older Adults and Parkinson’s Disease Patients. doi:10.21203/rs.3.rs-74250/v1.
- Sabo, A., Mehdizadeh, S., Ng, K., Iaboni, A., Taati, B., 2020. Assessment of parkinsonian gait in older adults with dementia via human pose tracking in video data. Journal of NeuroEngineering and Rehabilitation 17. doi:10.1186/s12984-020-00728-9.
- Sahoo, S., Saboo, M., Pratihari, D.K., Mukhopadhyay, S., 2020. Real-time detection of actual and early gait events during level-ground and ramp walking. IEEE Sensors Journal 20, 8128–8136. doi:10.1109/JSEN.2020.2980863.
- Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y., 2016. Geinet: View-invariant gait recognition using a convolutional neural network, pp. 1–8. doi:10.1109/ICB.2016.7550060.
- Sikandar, T., Rabbi, M.F., Ghazali, K.H., Altwijri, O., Alqahtani, M., Almijalli, M., Altayyar, S., Ahamed, N.U., 2021. Using a deep learning method and data from two-dimensional (2d) marker-less video-based images for walking speed classification. Sensors 21, 2836. doi:10.3390/s21082836.
- Simonetti, E., Villa, C., Bascou, J., Vannozzi, G., Bergamini, E., Pillet, H., 2019. Gait event detection using inertial measurement units in people with transfemoral amputation: a comparative study. Medical Biological Engineering Computing 58. doi:10.1007/s11517-019-02098-4.
- Singh, J.P., Jain, S., Arora, S., Singh, U.P., 2018. Vision-based gait recognition: A survey. IEEE Access 6, 70497–70527. doi:10.1109/ACCESS.2018.2879896.
- Tang, Y., Li, Z., Tian, H., Ding, J., Lin, B., 2019. Detecting toe-off events utilizing a vision-based method. Entropy 21. doi:10.3390/e21040329.
- Verlekar, T., Correia, P., Soares, L., 2017. View-invariant gait recognition system using a gait energy image decomposition method. IET Biometrics 6. doi:10.1049/iet-bmt.2016.0118.
- Verlekar, T., Soares, L., Correia, P., 2018. Automatic classification of gait impairments using a markerless 2d video-based system. Sensors 18, 2743. doi:10.3390/s18092743.
- Verlekar, T.T., Vroey, H.D., Claeys, K., Hallez, H., Soares, L.D., Correia, P.L., 2019. Estimation and validation of temporal gait features using a markerless 2d video system. Computer Methods and Programs in Biomedicine 175, 45–51. URL: <https://www.sciencedirect.com/science/article/pii/S016926071831040X>, doi:<https://doi.org/10.1016/j.cmpb.2019.04.002>.
- Voeikov, R., Falaleev, N., Baikulov, R., 2020. TNet: Real-time temporal and spatial video analysis of table tennis.
- Xu, X., McGorry, R., Chou, L.S., Lin, J.H., Chang, C.C., 2015. Accuracy of the microsoft kinecttm for measuring gait parameters during treadmill walking. Gait Posture 34. doi:10.1016/j.gaitpost.2015.05.002.
- Yang, C., Ugbohue, U., Kerr, A., Stankovic, V., Stankovic, L., Carse, B., Kaliarntas, K., Rowe, P., 2016. Autonomous gait event detection with portable single-camera gait kinematics analysis system. Journal of Sensors 2016, 1–8. doi:10.1155/2016/5036857.
- Zahradka, N., Verma, K., Behboodi, A., Bodt, B., Wright, H., Lee, S., 2020. An evaluation of three kinematic methods for gait event detection compared to the kinetic-based ‘gold standard’. Sensors 20, 5272. doi:10.3390/s20185272.
- Zhang, S., Wang, Y., Li, A., 2021. Cross-view gait recognition with deep universal linear embeddings, pp. 9091–9100. doi:10.1109/CVPR46437.2021.00898.