

Disease Estimation Using Gait Videos by Separating Individual Features Based on Disentangled Representation Learning

Shiori Furukawa^a and Noriko Takemura^b

Kyushu Institute of Technology, Fukuoka, Japan

Keywords: Gait, Disease-Estimation, Image-Processing.

Abstract: With the aging of society, the number of patients with gait disturbance is increasing. Lumbar spinal canal stenosis (LCS) and cervical spondylotic myelopathy (CSM) are representative diseases that cause gait disturbance. However, diagnosing these diseases takes a long time because of the wide variety of medical departments and lack of screening tests. In this study, we propose a method to recognize LCS and CSM using patients' walking videos. However, the gait images of patients contain not only disease features but also individual features, such as body shape and hairstyle. Such individual features may reduce the accuracy of disease estimation. Therefore, we aim to achieve highly accurate disease estimation by separating and removing individual features from disease features using a deep learning model based on a disentangled representation learning approach. In evaluation experiments, we confirmed the usefulness of the proposed method by verifying the accuracy of different model structures and different diagnostic tasks to be estimated.

1 INTRODUCTION

Gait disturbance is one of the most common disorders in an aging society. Gait disturbance not only restricts the patient's activities but also has psychological effects, such as memory loss and decreased motivation caused by decreased walking time. From a social point of view, it is also a problem that it requires much effort to care for patients with gait disturbance.


Typical diseases with gait disturbance include lumbar spinal canal stenosis (LCS), cervical spondylotic myelopathy (CSM), Parkinson's disease, peripheral arterial disease, and cerebrovascular disease. Because of the wide variety of departments specializing in these diseases and the lack of simple screening tests, such as biomarkers, it can take considerable time to receive a correct diagnosis; 43% of patients with cervical spondylosis are initially diagnosed with other diseases and patients with gait disorders visit 5.2 physicians on average before receiving an appropriate diagnosis (Wu et al., 2013). In this study, we aim to automatically estimate these diseases based on a person's gait characteristics.


Several studies have been conducted on gait anal-

ysis for diseases with gait disorders (Abdulhay et al., 2018; Tahir and Manap, 2012; Kidziński et al., 2020; Nguyen et al., 2016). Tahir et al. (Tahir and Manap, 2012) used a motion capture system and floor reaction force meter to extract features such as the joint angle, stride length, and floor reaction force during walking, and used a machine learning model to identify patients with Parkinson's disease. However, this method uses expensive sensors that require specialized knowledge, which makes it unsuitable for practical diagnosis and screening tests.

By contrast, Kidziński (Kidziński et al., 2020) estimated gait speed, cadence, the knee joint angle, and other parameters using gait videos captured by a single camera. Although this method is highly practical because gait features can be estimated simply by capturing a person walking using a camera, it estimates the above features based on a rough skeletal model. It lacks information closely related to diseases, such as a subtle bending of the neck and hips. Furthermore, estimation errors and false positives for the joint points may lead to a decrease in the accuracy of disease estimation. Therefore, in this study, we adopt an appearance-based method with silhouette features instead of a model-based method with skeletal features to estimate diseases from gait videos.

Appearance-based methods directly estimate a disease from images; hence, little information about

^a  <https://orcid.org/0009-0008-4614-6722>

^b  <https://orcid.org/0000-0003-1977-4690>

the disease is missing. However, simultaneously, personal characteristics, such as hairstyle and body shape, are also included in the images, and these may affect the performance of disease estimation. In this study, we address this problem using disentangled representation learning (DRL), which can separate features. DRL is often used to generate face images in which only facial expressions and poses are changed (Tran et al., 2017; Higgins et al., 2016). In this study, we apply the DRL framework used as an image generator as a discriminator. As a DRL model, the variational autoencoder (VAE) (Kingma and Welling, 2013) is often used. However, a VAE-based model includes a decoder for reconstructing images, which is not necessary for the classification tasks (Shiori Furukawa, 2024). In this study, we aim to improve accuracy by modifying the network to a convolutional neural network (CNN), which is used as a feature extractor (Donahue et al., 2014) and specialized for classification tasks, and comparing it to VAE. Using 263 people’s walking videos, LCS, CSM, and healthy discrimination were analyzed to confirm the usefulness of the proposed method.

2 PROPOSED METHOD

In this study, we estimate diseases using a mean silhouette image (gait energy image, GEI (Han and Bhanu, 2005)) generated from walking videos. We aim to improve accuracy using a VAE-based DRL model and a CNN-based DRL model to separate disease features and individual features. The details of the proposed method are described below.

2.1 Gait Features

Silhouettes are extracted from walking videos and an average silhouette image normalized by height: GEI (128×88 pixels) is generated. A graph transition (Gong et al., 2019) is used for the person region segmentation method. Because patients with gait disorders have unstable gait cycles, the number of frames used to generate the GEI was experimentally set to 40 frames.

GEI is a practical gait feature used in various studies on gait analysis and recognition because it represents static features, such as neck and back flexion, and dynamic features, such as limb swing, in a single image (Sakata et al., 2019; Takemura et al., 2018; Liao et al., 2021). By contrast, as shown in Figure 1, it also includes many individual features, such as body shape and hairstyle; hence, it is necessary to consider the effects of such individual differences when ana-



Figure 1: These GEIs all belong to different individuals. GEIs include many individual features, such as body shape and hairstyle; hence, it is necessary to consider the effects of such individual differences when analyzing gait.

lyzing gait.

2.2 Disease Estimation Method Using a VAE

2.2.1 DRL Model

In this study, we perform feature separation in the latent space based on Guided-VAE (Ding et al., 2020). The VAE-based DRL model consists of three networks: a VAE model, an excitation classifier, and an inhibition classifier, as shown in Figure 2. The details of each network structure are the same as those in (Ding et al., 2020).

VAE Model. The network reconstructs the same image as the input image after compressing the input image once. The loss function L_{VAE} (Equation 3) is the sum of the reconstruction error (mean squared error, Equation 1) of the input and output images, and the Kullback-Leibler divergence (KLD, Equation 2) measures the difference between two probability distributions. In the context of Variational Autoencoders (VAE), we compare the latent variable distribution as $Q(z|x)$ with the prior distribution as $P(z)$, which is typically assumed to be a standard normal distribution.

$$L_{recon} = \frac{1}{N} \sum_{i=1}^N \|x_i - \hat{x}_i\|^2 \quad (1)$$

$$L_{KLD(P|Q)} = \sum_z Q(z|x) \log \frac{Q(z|x)}{P(z)} \quad (2)$$

$$L_{VAE} = L_{recon} + L_{KLD} \quad (3)$$

Excitation Classifier. The classifier is used when learning so that a latent variable obtains specific feature information, where the loss function L_{exc} is Binary Cross-Entropy Loss (BCE) or Cross-Entropy Loss (CE).

Inhibition Classifier. The classifier is used when learning, so that the remaining latent variables do not have specific feature information, and the loss function L_{inh} is BCE or CE, as is L_{exc} .

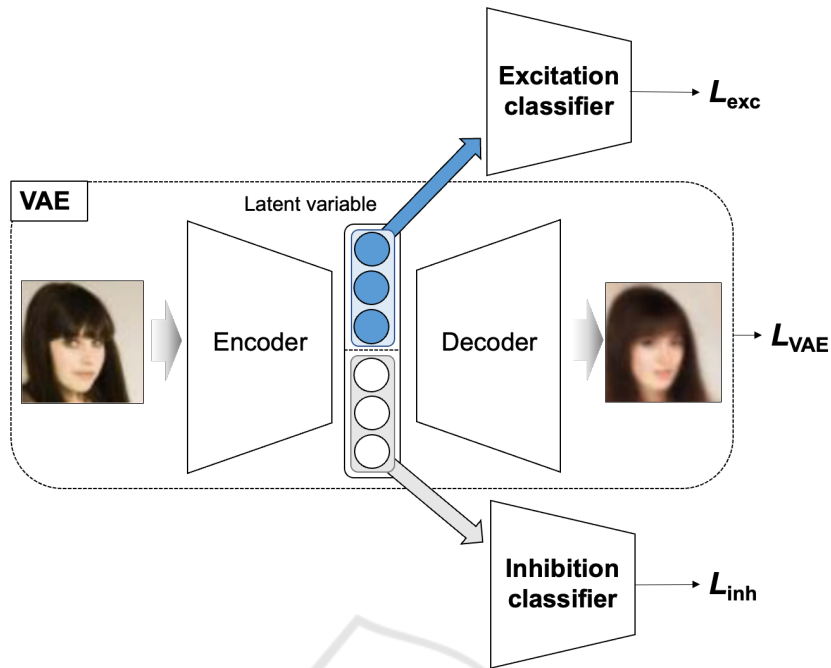


Figure 2: In this study, we perform feature separation in the latent space based on Guided-VAE (Ding et al., 2020). The VAE-based DRL model consists of three networks: a VAE model, an excitation classifier, and an inhibition classifier. The details of each network structure in our proposed method are the same as this DRL model.

The learning procedure for this VAE-based DRL model is shown below. Repeating this learning procedure can separate latent variables with and without specific feature information. In the proposed method, the procedures are applied to both disease features and individual features.

[Learning Procedure of VAE-Based DRL Model]

- (1) The VAE and excitation classifier parameters are trained with the loss function as $L_{VAE} + \alpha L_{exc}$, where α is the weight of the sum of L_{VAE} (Equation 3) and L_{exc} . These are learned so that the excitation classifier can classify feature labels correctly, the VAE can reconstruct the correct image, and the latent variables contain features that the excitation classifier can classify correctly.
- (2) An inhibition classifier is trained with the loss function as L_{inh} . The inhibition classifier is trained to classify feature labels correctly.
- (3) The feature label (one-hot vector) is set to uniform values (label value = $1/\#classes$) and the VAE is trained with the loss function as L_{inh} . Latent variables are trained so they do not have specific feature information.

Disease estimation is performed considering individual features using the VAE-based DRL model indicated above. The framework of the method is shown in Figure 3. The input GEI for model training is la-

beled with the presence or absence of a disease and an ID. In the proposed method, the above procedure is repeated for both diseases and individuals. When learning individual features, the latent variables used as inputs for the excitation classifier and the inhibition classifier are reversed compared to when learning disease features. Eventually, the excitation classifier for disease features is used to estimate diseases. In this way, disease estimation that accounts for individual differences can be performed using latent variables that capture disease features but exclude individual features.

2.3 Disease Estimation Method Using a CNN

Latent variables in the VAE include features that can reconstruct the image, i.e., all features related to the image. Therefore, it is necessary to separate the latent variables into parts that have specific features and parts that do not. By contrast, latent variables in the CNN extract only features related to the specific feature; hence, there is no need to separate latent variables unrelated to the specific feature. The CNN model does not have the task of reconstructing the image, which allows it to focus more on the classification task. Therefore, the learning procedure of the CNN excludes some steps from the VAE learning

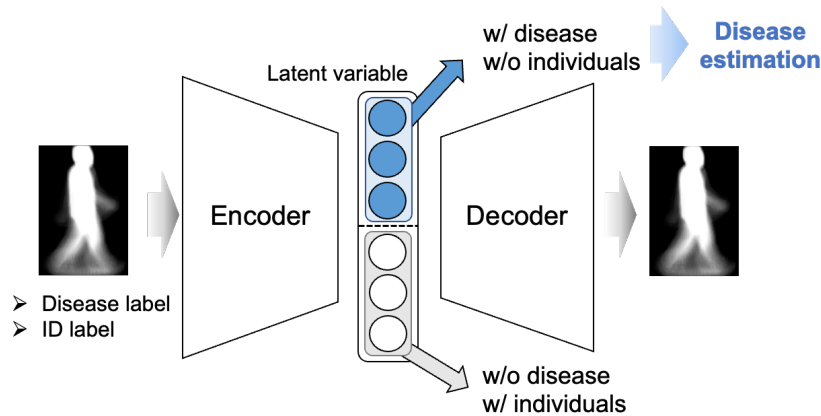


Figure 3: This is the framework of the proposed method utilizing a VAE. By applying [Learning Procedure of VAE-based DRL model] with disease features, the latent variables are divided into those representing disease (*w/ disease*) and those not representing disease (*w/o disease*). Similarly, by applying [Learning Procedure of VAE-based DRL model] with individual features, the latent variables are divided into those excluding individual-specific features (*w/o individuals*) and those including individual-specific features (*w/ individuals*).

procedure. The framework of the method is shown in Figure 4. The CNN network shares the same architecture as the encoder in the VAE and the loss function L_{KLD} represents the Kullback-Leibler divergence. Unlike the VAE, the CNN does not perform image reconstruction; thus, reconstruction error is not included. However, L_{KLD} , which encourages dimensional independence in the latent variable space, is utilized.

The learning procedure for this CNN-based DRL model is described below. Step (1) is applied to disease features to enable disease identification, while step (2) and step (3) are applied to individual features to ensure individuals cannot be identified.

[Learning Procedure of CNN-Based DRL Model]

- (1) The CNN and excitation classifier parameters are trained with the loss function as $L_{KLD} + \beta L_{exc}$, where β is the weight of the sum of L_{KLD} (Equation 2) and L_{exc} . These are learned so that the excitation classifier can classify disease labels correctly, i.e., the latent variables contain features that diseases can be classified correctly.
- (2) An inhibition classifier is trained with the loss function as L_{inh} . The inhibition classifier is trained to classify ID labels correctly.
- (3) The feature label (one-hot vector) is set to uniform values (label value = $1/\#ID$) and the CNN is trained with the loss function as L_{inh} . Latent variables are trained so they do not have individual features.

3 EVALUATION

For performance evaluation, we collected gait videos of patients with gait disorders and normal subjects.

3.1 Dataset

A standard monocular RGB camera captured 4 meters of the distance the people walked of LCS patients, CSM patients, and healthy subjects (1288×964 pixels, 30 fps). Figure 5 shows an example of the captured gait videos. A total of 139 LCS patients, 59 CSM patients (19 of whom had both LCS and CSM), and 84 healthy subjects had gait videos collected one to four times per person, for a total of 896 times. A physician's diagnosis disease labels and ID labels were assigned to each gait video. GEIs were generated from 40 frames of image sequences by staggering the images by one frame. The number of subjects and the number of GEIs are shown in Table 1. Note that the GEIs used for model training had an upper limit of 200 images per video and were undersampled to eliminate data bias.

3.2 Comparison Methods

To demonstrate the usefulness of the proposed disease estimation method that considers individual differences, we evaluated its performance using the following comparative methods.

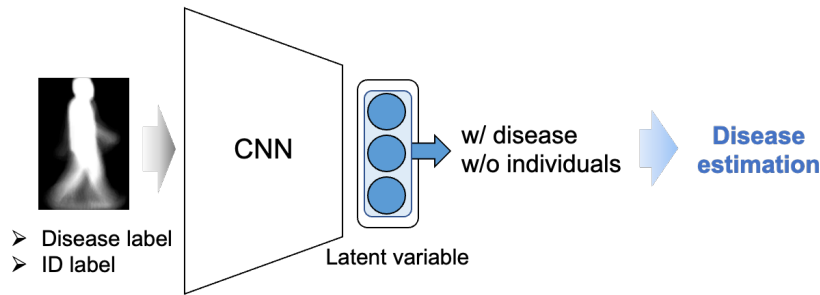


Figure 4: This is the framework of the proposed method utilizing a CNN. In the **[Learning Procedure of CNN-based DRL model]**, step (1) is applied to disease features to enable disease identification (*w/* disease), while step (2) and step (3) are applied to individual features to ensure individuals cannot be identified (*w/o* individuals).

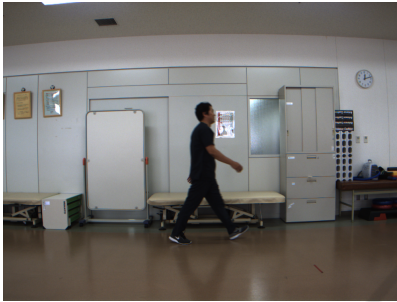


Figure 5: This is an example of a gait video. A standard monocular RGB camera captured 4 meters of the distance the people walked of LCS patients, CSM patients, and healthy subjects (1288×964 pixels, 30 fps).

3.2.1 VAE-Based Model

Comparison Method 1 (Comp1). Only the **[Learning Procedure of VAE-Based DRL Model]** with disease features is performed; those of individual features is not performed.

Comparison Method 2 (Comp2). Only the **[Learning Procedure of VAE-Based DRL Model]** step (1) with disease features is performed. All latent variables were trained to identify the disease using a disease excitation classifier and VAE. However, to keep the latent variable dimension used for disease estimation the same as that for the other methods, the number of latent variables is half that of the other methods.

3.2.2 CNN-Based Model

Comparison Method 3 (Comp3). Only the **[Learning Procedure of CNN-Based DRL Model]** step (1) with disease features is performed. All latent variables are trained to identify the disease using a disease excitation classifier and CNN.

Table 1: The number of subjects and the number of GEIs are shown in the below table. Note that the GEIs used for model training had an upper limit of 200 images per video and were undersampled to eliminate data bias.

	#Subjects	#Movies	#GEIs
LCS	139	504	48,456
CSM	59	192	26,970
healthy	84	269	28,609

3.3 Evaluation Method

Using gait data from LCS patients, CSM patients, and healthy subjects, we generated the following four disease estimators and evaluated the performance of the proposed and comparison methods, respectively.

LCS Estimator: LCS vs. {CSM, healthy}

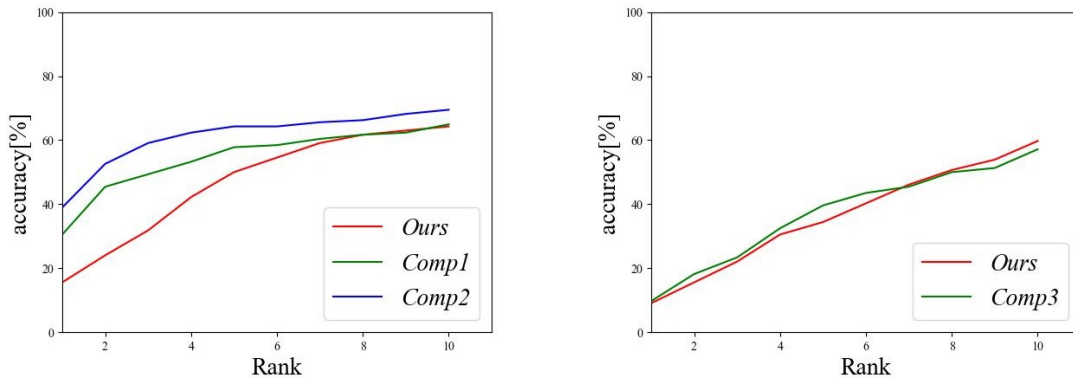
CSM Estimator: CSM vs. {LCS, healthy}

Disease Estimator: {LCS, CSM} vs. healthy

Multi Class Estimator: LCS vs. CSM vs. healthy

For the performance evaluation, we divided videos into ten groups and conducted 10-fold cross-validation: one group was the test data, another was the validation data, and the remaining eight were the training data. The average F1 score of the 10 groups was used for the evaluation.

The mini-batch size for training was set to 512 and the model's performance was evaluated on the test data using up to 100 epochs, selecting the epoch where the F1 score of the validation data was maximized. Hyperparameters such as L_{VAE} and L_{exc} weights α , L_{KLD} and L_{exc} weights β , the initial learning rate, and weight decay for each network were determined using the validation data at the first cross-validation, and the same values were used for the second and subsequent cross-validations. Adam was used for model optimization (Kingma and Ba, 2014).



(a) VAE-based model

(b) CNN-based model

Figure 6: Rank- N accuracy for individual identification.

3.4 Result

The results of the evaluation experiment are shown in Table 2. For all estimators except Dis_{est} of the CNN, the proposed method was more accurate than the comparison methods. Therefore, we demonstrated the effectiveness of the proposed method for separating individual features in the latent space.

By contrast, $Comp3$ with the CNN obtained the highest accuracy for Dis_{est} . As their high average accuracies indicate, the task were easier than those for other estimators, and even a simple model achieved sufficiently high accuracy. Therefore, it seems that the disadvantages of complex models that make learning more difficult outweigh the advantages of considering individual differences in the proposed method.

The results of the VAE and CNN were compared. The proposed method with the CNN was more accurate than the proposed method with the VAE except for Dis_{est} . We demonstrated the effectiveness of a CNN specialized for classification tasks.

Individual identification was also analyzed to verify the extent to which individual features were removed from the latent variable used for disease estimation. Figure 6 shows the rank- N accuracy of individual identification. The intermediate layer output of the individual identification classifier was a feature vector and the feature vectors of all subjects in the dataset were obtained in advance as registration data. The L2 norm computed by the feature vector for a given input and the feature vector of each registered data were sorted in decreasing order and the proportion of the same person in the top N subjects was calculated (Phillips et al., 2000). The smaller the proportion, the less identifiable the individual, that is, the more separated the individual features. As Figure 6 shows, the proposed method separated individual features better than the comparison methods.

Table 2: Average F1 score for each estimator that LCS estimator (LCS_{est}), CSM estimator (CSM_{est}), Disease estimator (Dis_{est}), and Multi class estimator (Mul_{est}). Bold indicates the best value.

(a) VAE-based model

	LCS_{est}	CSM_{est}	Dis_{est}	Mul_{est}
<i>Ours</i>	0.924	0.844	0.979	0.760
<i>Comp1</i>	0.916	0.843	0.971	0.744
<i>Comp2</i>	0.915	0.842	0.977	0.755

(b) CNN-based model

	LCS_{est}	CSM_{est}	Dis_{est}	Mul_{est}
<i>Ours</i>	0.925	0.859	0.978	0.772
<i>Comp3</i>	0.914	0.857	0.979	0.765

4 CONCLUSION

In this study, we proposed a method for disease estimation from gait videos. We aimed to improve disease estimation accuracy by separating disease and individual features in the latent space of a VAE and a CNN using the DRL model.

Almost all of the proposed methods were more accurate than the comparison methods, which demonstrates the effectiveness of the methods for separating disease and individual features. Additionally, almost all the proposed methods obtained better CNN accuracy than VAE, which indicates the effectiveness of the specialized model for the classification tasks proposed in this study.

We plan to expand the scope to include diseases other than LCS and CSM, such as Parkinson's. We only used the side video of walking in the experiments, but we aim to further improve accuracy by also using features from the frontal video. Additionally,

we aim to further improve accuracy using MRI and CT images in addition to gait images. We will verify the usefulness of this method for separating individual features for other tasks, such as facial expression recognition. In this study, we conducted experiments using VAE and CNN to verify the effectiveness of our feature separation method. Furthermore, since our feature separation method can be applied to various backbones, we plan to apply it to more tasks using existing networks.

REFERENCES

- Abdulhay, E., Arunkumar, N., Narasimhan, K., Vellaiappan, E., and Venkatraman, V. (2018). Gait and tremor investigation using machine learning techniques for the diagnosis of parkinson disease. *Future Generation Computer Systems*, 83:366–373.
- Ding, Z., Xu, Y., Xu, W., Parmar, G., Yang, Y., Welling, M., and Tu, Z. (2020). Guided variational autoencoder for disentanglement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7920–7929.
- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR.
- Gong, K., Gao, Y., Liang, X., Shen, X., Wang, M., and Lin, L. (2019). Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7450–7459.
- Han, J. and Bhanu, B. (2005). Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-vaе: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.
- Kidziński, Ł., Yang, B., Hicks, J. L., Rajagopal, A., Delp, S. L., and Schwartz, M. H. (2020). Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature communications*, 11(1):4054.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Liao, R., Moriwaki, K., Makihara, Y., Muramatsu, D., Takemura, N., and Yagi, Y. (2021). Health indicator estimation by video-based gait analysis. *IEICE TRANSACTIONS on Information and Systems*, 104(10):1678–1690.
- Nguyen, T.-N., Huynh, H.-H., and Meunier, J. (2016). Skeleton-based abnormal gait detection. *Sensors*, 16(11):1792.
- Phillips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J. (2000). The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104.
- Sakata, A., Takemura, N., and Yagi, Y. (2019). Gait-based age estimation using multi-stage convolutional neural network. *IPSJ Transactions on Computer Vision and Applications*, 11:1–10.
- Shiori Furukawa, N. T. (2024). Disease estimation based on gait images by separating individual features using variational autoencoder. In *AROB-ISBC-SWARM 2024*.
- Tahir, N. M. and Manap, H. H. (2012). Parkinson disease gait classification based on machine learning approach. *Journal of Applied Sciences(Faisalabad)*, 12(2):180–185.
- Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., and Yagi, Y. (2018). Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ transactions on Computer Vision and Applications*, 10:1–14.
- Tran, L., Yin, X., and Liu, X. (2017). Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Wu, J.-C., Ko, C.-C., Yen, Y.-S., Huang, W.-C., Chen, Y.-C., Liu, L., Tu, T.-H., Lo, S.-S., and Cheng, H. (2013). Epidemiology of cervical spondylotic myelopathy and its risk of causing spinal cord injury: a national cohort study. *Neurosurgical focus*, 35(1):E10.