MMU GAG Dataset: A Comprehensive Dataset for Gait Age and Gender Recognition

Tee Cheng Qi1, a), Tee Connie1, 2, b), Timilehin B. Aderinola3, c) and Michael Kah Ong Goh1, 2, d)

1Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

2Centre for Image and Vision Computing, CoE for Artificial Intelligence, Multimedia University, Jalan Ayer Keroh Lama, 75450, Melaka, Malaysia

3University College Dublin, School of Computer Science, Science Centre, East Belfield Dublin

b) Corresponding author: tee.connie@mmu.edu.my

*a) 1211203730@student.mmu.edu.my*

*c) timilehin.aderinola@ucd.ie*

*d) michael.goh@mmu.edu.my*

**Abstract.** Gait recognition has become an important biometric recognition technology due to its advantages of non-contact and remote recognition. The fast development of industries such as personalized services has made age and gender estimation based on gait recognition emerge as a significant research topic. It can complete real-time analysis of crowd structure without disturbing the detected object. However, there is still a lack of gait datasets specifically for age and gender recognition in this field. So, this paper proposes and constructs the MMU GAG Gait Dataset. The dataset also contains two datasets. The first is a self-collected dataset containing 930 subjects, which strictly adheres to structured shooting. The second is the more challenging Gait in the Wild Dataset, which has 419 subjects involved. Unlike the self-collected dataset, this dataset has no restrictions on the number and direction of walking. These datasets demonstrate various crowd backgrounds in different shooting locations which strengthens the ability of model to perform in real-life scenarios. The consistency and reliability between datasets were verified by statistical analysis of key points. The paper offers novel ground for deploying gait features into real-time demographic assessment and personalized solutions.

# Introduction

In recent years, gait recognition has become one of the important research directions in the field of biometrics. Gait refers to the posture and behavioral characteristics of a person when walking [1]. By analyzing the gait and walking pattern characteristics of the human body, individuals can be effectively identified or distinguished. This technology has the advantages of non-invasive and long-distance recognition, and identity authentication can be completed without the active cooperation of the target. Therefore, compared with traditional biometric technologies such as fingerprints and irises that require close-range scanning, gait recognition is more applicable in public places.

In addition to identity authentication, some neurological diseases or movement disorders can also be analyzed through individual gait characteristics. Besides, human age and gender can also be inferred. This basic demographic information serves many practical needs for behavior analysis as well as personalized services and security monitoring purposes. If applied properly, this will bring many conveniences to human society.

Given the wide application of gait recognition in many fields, a complete dataset with high standardization and diversity is needed to support in-depth exploration in related fields. Therefore, this study focuses on building and using a large-scale gait dataset called MMU GAG Dataset for age and gender estimation that includes the Gait in the Wild dataset that reflects real scene changes and the dataset collected by the research team. The Gait in the Wild Dataset contains 419 subjects, and all are collected from many public data resources, so the number and direction of walking are not controlled. The dataset collected by the team has up to 930 subjects, and all video data are structured and processed according to unified standards to ensure the accuracy and consistency of the data. Therefore, compared with the Gait in the Wild subset, Self-Collected Dataset is more strictly controlled in terms of lighting conditions, shooting angles, background interference and other factors, and therefore has higher reliability and research value. Based on the above two subsets, this paper will conduct comparative experimental analysis around the gender and age estimation tasks, evaluate the performance differences of different gait features in various scenarios, and verify the feasibility and effectiveness of the MMU GAG dataset in real applications, so as to provide a solid data foundation and theoretical support for subsequent gait recognition research.

# literature review

Many studies investigating human gait cycle data through human pose estimation, gait energy images and inertial sensors have conducted numerous studies on age and gender analysis to identify distinct physical features since recent years. In terms of classification, gender is divided into male and female. Age classification is usually divided into three groups, namely children from 0 to 14 years old, adults from 15 to 64, and seniors of 65 or above.

Authors in [2] proposed that the OU-ISIR Gait Database, Multi-View Large Population Dataset. It is one of the databases in the OU-ISIR Gait Database, which was collected by Institute of Scientific and Industrial Research, Osaka University at the Science Museum and released on February 22, 2018. The database has 10307 different subjects aged from 2 to 87, including 5114 males and 5193 females. It also contains gait data shot at up to 14 angles, ranging from 0° to 90° and 180° to 270°, mainly used for gait recognition and analysis research.

Authors in [3] proposed the OU-ISIR Gait Database, Large Population Dataset with Age that collected by Osaka University Institute of Information Science at the Science Museum and released on December 22, 2017. The database contains 63846 subjects, ranging in age from 2 to 90 years old. The entire dataset is divided into training and test sets, each with 31923 subjects. In addition, the dataset provides a size-standardized GEI for each subject.

Authors in [4] proposed that the OU-ISIR Gait Database, Inertial Sensor Dataset. It is one of the databases in OU-ISIR Gait Database, collected by Osaka University Scientific and Industrial Research Center at the Science Museum and released on October 15, 2013. Data collection was performed using inertial sensors. The database has two subsets, the first contains 744 subjects aged from 2 to 78 years old, each with two horizontal walking sequences, 389 of whom are male and the remaining 355 are female. The second one has 903 subjects, each also has two horizontal walking, uphill and downhill sequences. This database is mainly used for gait recognition and analysis and is particularly suitable for gait data collected by mobile devices or wearable devices.

Authors [5] proposed that the OU-ISIR Large Population Gait Database with Real Carried Objects (OU-LP-Bag), which is one of the databases in the OU-ISIR gait database, collected by Osaka University Scientific and Industrial Research Center at the Science Museum and released on June 15, 2018. The dataset contains 62,528 subjects ranging in age from 2 to 95 years old, who all walked three times along a straight line at their preferred speed, and there will be corresponding (CS) labels according to different carrying states. The database contains real gait data of many subjects carrying different objects, including background-subtracted image sequences and related GEIs. The database is mainly used to study gait changes when carrying objects, simulating the gait characteristics of people carrying bags, backpacks and other objects in real life.

The OU-ISIR Gait Database [6], Similar Action Inertial Dataset is one of the databases in the OU-ISIR gait database, collected by the Industrial Science Research Center of the Science Museum of Osaka University and released on September 11, 2015. The dataset contains more than 460 subjects ranging in age from 8 to 78 years old. The database contains 5 types of gait actions, namely walking on flat ground, going up and down stairs, and going up and down hills. This dataset is designed to analyze gait changes in similar actions using inertial sensor data.

Zhang et al. (2022) [7] proposed that the Front-View Gait Blurred (FVG-B) Database. It is a gait database that focuses on gait data blurred from the frontal perspective. There are 226 subjects, which 146 are male and 90 are female. Since the dataset does not provide age and gender information, only gender information is annotated. The database is captured in near-frontal views (-45◦, 0◦, and 45◦) in different ways, and is intentionally blurred during the acquisition process to simulate the blur effect caused by camera movement, poor lighting, and other factors in real scenes.

Authors in [8] proposed that the Multimedia University Gait Age and Gender Dataset (MMU GAG), gait in the wild (GITW). The database contains gait data classified by age and gender, with a total of 439 subjects, each of whom walks at a normal speed in an unconstrained environment to obtain videos. In addition, each subject is labeled with their respective gender and age group, including children, adults, or elderly.

Authors in [9] proposed that the USF HumanID Gait Baseline Database which is collected at the University of South Florida (USF) on May 20 & 21 and November 15 & 16, 2001. The database has 122 subjects aged 19 to 59 years, including 91 males and 31 females. The database has 32 sequences consisting of gaits in all possible combinations of up to five factors. These conditions include two camera angles and shoe types, grass and concrete surfaces, carrying and not carrying a briefcase, and two different dates six months apart. Table 1 shows that the summary of all gait datasets.

TABLE 1. A summary of gait datasets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Datasets** | **Input Devices** | **Classes** | **Number of Sample** | **Subject** |
| The OU-ISIR Gait Database, Multi-View Large Population Dataset | Camera-Based | Gender + Age | 288596 sequences | 10307 |
| The OU-ISIR Gait Database, Large Population Dataset with Age | Camera-Based | Gender + Age | 63846 sequences | 63846 |
| The OU-ISIR Gait Database, Inertial Sensor Dataset | Inertial Sensor-Based | Gender + Age | 5100 sequences | 1647 |
| The OU-ISIR Large Population Gait Database with real-life carried object | Camera-Based | Gender + Age | 187584 sequences | 62528 |
| The OU-ISIR Gait Database, Similar Action Inertial Dataset | Inertial Sensor-Based | Gender + Age | 2300 gait actions | 460 |
| The Front-View Gait Blurred Database | Camera-Based | Gender | 2856 videos | 226 |
| Multimedia University Gait Age & Gender Datasets | Camera-Based | Gender | 439 json | 439 |
| USF HumanID Gait Baseline database | Camera-Based | Gender + Age | 1870 sequences | 122 |

# MMU GAG Dataset

# Self collected dataset

In MMU GAG Dataset, the Self-Collected Dataset collected 930 complete gait videos. All subjects must be set up and photographed according to the designated walking path (Figure 1). During the entire filming process, the camera needs to be fixed in front of each subject. Then, each subject will walk to the left or right at least 3 times. After completion, they need to move to the center point of the path and change the direction to walk straight forward and walk straight backward at least 3 times. Each video strictly follows the set walking path to obtain gait information from the relevant video.

In the initial stage of gait video collection, in order to make the subjects feel at ease to participate, the subjects will be asked to fill in the consent form. After that, several sample videos will be provided as a demonstration to let them understand how to shoot the required correct gait videos. After receiving the video, each collected video will be carefully reviewed to ensure the integrity of the video. If there are obvious errors such as missing walking times, the video is too blurry, etc., the subject will be asked to re-shoot the correct gait video if it is convenient. Below show that the video samples of Self-Collected Dataset (Figure 2).

A diagram of a camera and subject

Description automatically generated

**FIGURE 1.** The correct walking path

## GAIT IN THE WILD DATASET

For Gait in the Wild Dataset, a total of 419 videos were collected. It is collected from many public data resources such as Weibo, Bilibili, Douyin, YouTube, etc. Unlike the structured shooting MMU GAG dataset, this dataset aims to increase the challenge of the task, so there is no restriction on the number and direction of walking. Real-world gait behavior in natural environments lies closer to the content of these videos which makes these datasets better than structured datasets for testing gait recognition models' capacity to work in realistic conditions. Below shows that the video sample of Gait in the Wild Dataset (Figure 3).

|  |  |
| --- | --- |
|  |  |
| (a) Front-back View | (b) Left-right View |

**FIGURE 2.** Video samples of self-collected dataset

**A person in shorts and a white shirt

AI-generated content may be incorrect.**

**FIGURE 3.** Video sample of gait in the wild dataset

## Demographic Information and Statistics

After implementing the Human Pose Tracking methodology to extract human body keypoints, a self-collected dataset contains 960 subjects and a total of 3963 JSON files. There are 340 males and 370 females, and the ethnicity includes Chinese, Malay, and Indian. In addition, these videos were shot both indoors and outdoors.

The Public dataset that contains 419 subjects and JSON files. The subjects consisted of 239 men and 180 females, each with only one gait sample JSON files. Since the public datasets are collected from online video, some of them involve individuals moving relative to the camera. The movement of camera subjects leads to view changes along with hidden objects and motion-based blur effects, which reduces the reliability of the dataset for robust classification. Each dataset has 136 keypoints, but some of the data from Self-Collected Dataset only have 17 keypoints.

The details of all datasets are shown in Table 2.

TABLE 2. A summary of datasets

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Keypoint** | **Subject** | **Child** | | **Adult** | | **Senior** | | **Sample** |
| **Male** | **Female** | **Male** | **Female** | **Male** | **Female** |
| Gait in the Wild | 136 | 419 | 82 | 67 | 100 | 71 | 57 | 42 | 419 |
| Self-Collected | 17/136 | 930 | 110 | 109 | 251 | 279 | 37 | 44 | 3963 |

All datasets show a wide range of values ​​in both the x and y coordinates, with the maximum value of the x coordinate reaching 3840.26 and the maximum value of the y coordinate reaching 3824.24. Gait in the Wild Dataset show a maximum standard deviation of 388.2, reflecting more significant position fluctuations. Furthermore, the median values of all datasets maintain parallelism with mean data points which reflects symmetrical distribution patterns. Table 3 analyzes the coordinate distribution patterns and differences in each dataset.

TABLE 3. Distribution patterns for each dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Coordinate** | **Value** | | | | |
| **Maximum** | **Minimum** | **Median** | **Mean** | **Standard Division** |
| Gait in the Wild | x | 3840.26 | -54.13 | 520.96 | 561.5 | 388.2 |
| y | 2185.62 | -21.23 | 410.57 | 450.6 | 230.71 |
| Self-Collected | x | 3833.81 | -30.91 | 915.1 | 874.13 | 346.06 |
| y | 3824.24 | -105.67 | 529.22 | 529.46 | 182.47 |

The MMU GAG Dataset stands out from current databases because it offers the following advantages.

1. **Data Diversity:** In order to enhance the breadth and challenge of research, this study designed and used two different types of gait datasets. Although the Self-Collected dataset is rich in data, it is in a controlled environment. So, we additionally prepared Gait in the Wild dataset.
2. **Gender balance:** Male subjects make up about 48% and females make up about 52% within the Self-Collected Dataset while Gait in the Wild Dataset shows about 57% male subjects along with about 43% female subjects. Overall, it still helps to achieve fairer and less biased model performance.
3. **Quality control:** Each keypoint sequence is manually visually inspected to remove erroneous key points caused by background interference, etc., to ensure the reliability of posture information.

# Experimental resultS

In this study, the Arsenal classifier was used to test the classification performance of each sub-dataset in the MMU GAG Dataset and analyze its accuracy in the gender and age recognition tasks. For the age estimation of Self-Collected Dataset, the model performed best in predicting the "Adult" category, correctly predicting 646 out of 729 samples, it achieves precision rate of 0.76, recall rate of 0.89 and F1 score of 0.82. For the "Child" category, the model correctly predicts 208 out of 341 samples, which get the precision rate of 0.72, recall rate of 0.61 and F1 score of 0.66. However, model shows weak performance in the "Senior" category which predicted correctly 34 out of 119 samples, which get the precision rate of 0.76, recall rate of 0.29 and F1 score of 0.41.

For the Gait in the Wild Dataset, the model still performs well in identifying "Child", with 38 samples correctly predicted out of 48. It achieves precision and F1 score of 0.78 and recall of 0.79. Besides, the "Adult" category is predicted correctly with 33 out of 51, which get the precision of 0.66, recall and F1 score of 0.65. For the "Senior" category, it is still the most challenging, only 12 samples correctly identified out of 27. Its precision, recall and F1 score are all 0.44, reflecting that still faces certain difficulties in identifying the elderly in real complex environments. For the gender estimation of Self-Collected dataset, the model performed well in identifying "Male", correctly predicts 406 out of 581 samples. For the "Female" category, 440 out of 608 samples were correctly identified. In terms of specific indicators, the precision rate for males was 0.71 and that for females was 0.72. The recall rate was 0.70 for males and 0.72 for females, and the F1 score was 0.70 for males and 0.72 for females. This shows that the model has good recognition effect and balance between categories in gender classification tasks.

For the Gait in the Wild dataset, the model still performs well in identifying "Male", correctly predicts 54 out of 66 samples. However, the recognition of "Female" is obviously insufficient, with only 19 out of 60 samples correctly predicted. This is showing that the model's ability to identify females has declined when dealing with more challenging real-world scenarios. In this dataset, the precision rate of males is 0.57, the recall rate is 0.82, and the F1 score is 0.67. The precision rate of females is 0.61, the recall rate is only 0.32, and the F1 score is 0.42. Although the model performs weakly in the female category, the overall performance remains at a high level in the male category. From Table 4, we can see that the model's age estimation accuracy in the self-collected data set is 74.68% and gender is 71.15%. This shows that the model has strong stability and reliability in a controlled environment. Age estimation accuracy in the Gait in the Wild dataset performed at only 65.87% and gender recognition accuracy reached a low of 57.94%. This shows that the model still faces high uncertainty in actual complex scenarios.

TABLE 4. Comparison of age and gender estimation

|  |  |  |
| --- | --- | --- |
| **Classification** | **Accuracy** | |
| **Self-Collected Dataset** | **Gait in the Wild Dataset** |
| Age | 74.68% | 65.87% |
| Gender | 71.15% | 57.94% |

## DISCUSSIONS

According to research, it can find that:

* Among each dataset, **age estimation has higher accuracy than gender estimation**. This may be because the gait patterns within different age groups appear more distinguishable. The identification of gender become harder due to the influence of posture and clothing because these external factors alter their characteristics.
* The age and gender **accuracy of the Gait in the Wild Dataset dataset is lower than the Self-Collected dataset**. Some of data in Gait in the Wild dataset involve individuals moving relative to the camera. The motion of camera subjects causes view variations together with hidden object and motion-based blurring effects that degrade the reliability of the dataset for robust classification.

# CONCLUSION

Our research aims to build the MMU GAG gait dataset which resolves the lack of suitable data for gait-based age and gender estimation while providing diverse and authentic gait samples. This dataset combines our self-collected data with the Gait in the Wild subset, significantly improving the generalization ability of the model in complex environments. The results show that Arsenal Classifier achieved the highest accuracy of 74.68% and 71.15% in the age & gender estimation tasks on the Self-Collected Dataset, respectively. It demonstrates its high efficiency in gait feature extraction under controlled conditions. On the more challenging Gait in the Wild dataset, the model achieves a 65.87% accuracy for age group categorization and 57.94% recognition rate for gender identification. Although complex environments place higher demands on the model, it can still maintain basic recognition capabilities and has a certain degree of generalization performance. In summary, this study developed a new age and gender estimation dataset for gait while proving the possibility of real-time population attribute identification through gait patterns. The identified advancements will boost applications of gait recognition systems throughout various fields. The model will gain additional practicality and adaptability as scene-diverse data gets added in the upcoming years.

# References

1. S.M.H.S.S. Fathima, and A. Valanarasi, “Human gait recognition using relevance vector machine classifier,” in *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, (IEEE, Ramanathapuram, India, 2016), pp. 564–568.
2. N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, “Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition,” IPSJ T Comput Vis Appl **10**(1), 4 (2018).
3. C. Xu, Y. Makihara, G. Ogi, X. Li, Y. Yagi, and J. Lu, “The OU-ISIR Gait Database comprising the Large Population Dataset with Age and performance evaluation of age estimation,” IPSJ T Comput Vis Appl **9**(1), 24 (2017).
4. T.T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, “The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication,” Pattern Recognition **47**(1), 228–237 (2014).
5. Y. Makihara, A. Suzuki, D. Muramatsu, X. Li, and Y. Yagi, “Joint Intensity and Spatial Metric Learning for Robust Gait Recognition,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, Honolulu, HI, 2017), pp. 6786–6796.
6. T.T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, and Y. Yagi, “Similar gait action recognition using an inertial sensor,” Pattern Recognition **48**(4), 1289–1301 (2015).
7. Z. Zhang, L. Tran, F. Liu, and X. Liu, “On Learning Disentangled Representations for Gait Recognition,” IEEE Trans. Pattern Anal. Mach. Intell. **44**(1), 345–360 (2022).
8. T.B. Aderinola, T. Connie, T.S. Ong, A.B.J. Teoh, and M.K.O. Goh, “Gait-based age group classification with adaptive Graph Neural Network,” Engineering Applications of Artificial Intelligence **137**, 109081 (2024).
9. S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, and K.W. Bowyer, “The humanID gait challenge problem: data sets, performance, and analysis,” IEEE Trans. Pattern Anal. Mach. Intell. **27**(2), 162–177 (2005).