Results and Discussion for Health&Gait: a video dataset for gait-based analysis

Jorge Zafra-Palma[1,2,*], Nuria Marín-Jiménez[3,4], José Castro-Piñero[3,4], Magdalena Cuenca-García[3,4], Rafael Muñoz-Salinas[1,2], and Manuel J.Marín-Jimenez[1,2,*]

[1]University of Cordoba, Department of Computing and Numerical Analysis, Córdoba, 14071, Spain

[2]Instituto Maimónides de Investigación Biomédica de Córdoba (IMIBIC), Córdoba, 14004, Spain

[3]GALENO Research Group, Department of Physical Education, Faculty of Education Sciences, University of Cadiz, Spain

[4]Instituto de Investigación e Innovación Biomédica de Cádiz (INiBICA), Cádiz, Spain

[*]Corresponding author(s): Jorge Zafra-Palma (jzafra@uco.es), Manuel J.Marín-Jiménez (mjmarin@uco.es)

In the following document, the results and discussions obtained from the experiments conducted using the dataset are presented in detail.

Experiments and Results

To demonstrate the quality control of the dataset, only state-of-the-art methods, which underwent a peer-review process, have been used to obtain the different provided data representations, as previously indicated in Sec. Methods (Subsec. Data processing). In addition, we have used the OptoGait system to measure gait parameters, which operates at a frequency of 1000 Hz and has an accuracy of 1 cm, as stated in its user manual (`https://medical.microgate.it/sites/default/files/manuali/optogait/Manual-EN.pdf`). The Optical Timing Gates (Photo-cells) from MuscleLAB, the system used to measure speed, have a resolution value of 2 ms, allowing for precise measurement of displacement times, as obtained from the MuscleLAB website
(`https://www.musclelabsystem.com/products/`). Furthermore, all measurements have been carried out under the supervision of sports science specialists.

Additionally, a series of computational experiments have been conducted to demonstrate the quality control of the dataset. For this, the experiments have been conducted to determine whether the biometric information of participants is encoded in the representations used for data. For this purpose, the task of sex classification and regression of participants' weight and age are used. For each task, the different data representations from the dataset videos, namely, silhouette, semantic segmentation, and optical flow, were used. Additionally, the gait parameters derived from the OptoGait and those estimated from the videos using pose information have been used to verify the utility of gait parameters in these tasks.

The tasks of sex classification and regression of age and weight are approached, on the one hand, using the Mobile Video Networks (MoViNets) [1] video classification architecture, training the model with information from the silhouette, semantic segmentation, and optical flow. On the other hand, a multilayer perceptron (MLP) and XGBoost [2] are trained using the gait parameters, which are *step length*, *stride length*, *cadence* and *speed*.

The MoViNets architecture has been selected as it is a Convolutional Neural Network (CNN) designed to achieve high classification rates while reducing computational and memory requirements, making it ideal for use on lower-performance devices, such as mobile devices. On the other hand, MLP and XGBoost have been chosen as they are state-of-the-art models in a wide variety of applications for classification and regression.

For the MoViNets architecture, the images need to be resized to 224×224 pixels and normalised. For the MLP and XGBoost models, an exhaustive hyperparameter tuning was conducted using the Bayesian Optimization algorithm [3]. The search space is the same for each of the experiments conducted. Below, the hyperparameter search space for multilayer perceptron is detailed.

- Number of layers: takes values in the range $[1, 5]$.
- Number of units per layer: takes values in the range $[32, 512]$ with a step of 32.
- Activation function: Activation function used in the hidden layers. Takes values $ReLU$ or $tanh$.
- Dropout: Boolean variable that indicates whether a dropout of 25% of the weights is performed or not.
- Learning rate: takes values in the range $[1 \times 10^{-4}, 1 \times 10^{-2}]$ using a logarithmic sampling.

Below, the hyperparameter search space for XGBoost is detailed.

- Max Depth: the maximum depth of a tree. Used to control over-fitting. Takes values in the range $[3, 10]$, and the values 12 or 15

- Min child weight: defines the minimum sum of weights of observations required in a child. Takes values in the range $[1, 8]$.
- Gamma: specifies the minimum loss reduction required to make a split. Takes values in the range $[0, 0.5]$ with a step of 0.1, and the values 1.0, 1.5 and 2.0.
- Subsample: denotes the fraction of observations to be random samples for each tree. Takes values in the range $[0.5, 1.0]$ with a step of 0.1.
- Colsample bytree: denotes the fraction of columns to be random samples for a tree. Takes values in the range $[0.5, 1.0]$ with a step of 0.1.
- Learning rate: takes values in the range $[0.01, 0.3]$ with a step of 0.05.
- Number of estimators: takes the value 50 or values in the range $[100, 1000]$ with a step of 100.
- Alpha: L1 regularization. Takes the values 0, 0.1, 3 or values in the range $[0.5, 2]$ with a step of 0.5.
- Lambda: L2 regularization. Takes the values in the range $[0.5, 2.0]$ with a step of 0.5, and the values 3, 4.5, 5, 6, 7 and 8.

To create the training and test sets, a 4-fold stratified split is applied based on the participants in the dataset, ensuring that there are no samples from the same participant in both the training and test sets. Additionally, 10% of the participants in the training set are selected for the validation set. The results obtained were averaged across each of the folds.

For the sex classification and age estimation tasks, participants are selected stratified according to their age and sex, ensuring a balanced representation of men and women in each age group. For the weight estimation task, participants are also selected in a stratified way, organised to ensure equal representation of men and women across the four body mass index (BMI) categories.

Moreover, all random number generators involved in obtaining the results have been seeded with a value of 27.

For classification, the average accuracy and F-Score across the four partitions were used, while for regression problems, the average mean absolute error along with the standard deviation.

Baselines for sex classification.

It is checked whether the information about the participant's sex is contained in the sequence of frames for each of the types of data present in the dataset, namely, silhouette, semantic segmentation, and optical flow, and it is also checked whether this information is also contained in the participant's gait parameters.

Regarding the data corresponding to silhouette, semantic segmentation, and optical flow, to check the influence of the different classes present in the dataset, the models are trained using samples from the class with a jacket, without a jacket, and combining both classes. To obtain the baselines, the MoViNet-A0-Base and MoViNet-A5-Base architectures were used to compare the influence of the model size on the final results. On the other hand, the *stream* version of the models (which is lighter in memory) was not chosen to try to obtain the best possible results. Both models have been pretrained on the Kinetics 600 dataset [4]. Fine-tuning was performed with the convolutional layer weights frozen while training two dense layers with 2,048 and two neurons, using the Swish activation function [5]. The diverse pieces of training were conducted over 50 epochs with early stopping set to a patience of 5, using a batch size of 32, and the weights were optimised using the Adam optimiser with a learning rate of 0.001. The number of epochs, batch size, and learning rate were selected through a hyperparameter tuning process using cross-validation on the validation data. Before obtaining the final results of the models, the influence of the number of selected frames on the obtained results is verified. To do this, the starting point is set as the MoViNet-A5-Base architecture using only data corresponding to WoJ class samples and silhouettes as the input data. Table 1 shows the results obtained by varying the length of the input frame sequence using silhouette as the data type. As the number of frames increases, accuracy also increases, although the increment becomes progressively smaller. Note that for the selection of frames, 15% of the frames at the beginning and the end of the sequence are excluded to ensure that the participants are fully present in the scene. A frame limit of 40 has been established, as there are clips in the dataset with only 40 selectable frames.

In Table 2, the results obtained for each method, for each data type and class, are shown. Based on the results presented in Table 1, the number of frames used was 40. It is observed that the best results were achieved using silhouette data for the MoViNet-A5-Base model trained on samples from both classes with an accuracy equal to 91.9%. The next best results were obtained using optical flow derived from the GMFlow learning model. It is observed that, except for semantic segmentation, the results obtained for the class WoJ are slightly superior, possibly because there may be some instances that, being covered up, are harder to classify. Additionally, the combination of samples from both classes, except again for semantic segmentation

Number of frames	Accuracy	F1 Macro
2	0.807	0.806
3	0.815	0.812
4	0.871	0.870
5	0.872	0.872
10	0.896	0.895
15	0.906	0.905
20	0.912	0.912
25	0.908	0.908
30	0.908	0.907
35	0.913	0.913
40	**0.918**	**0.918**

Table 1: Evaluation of the impact of frame number on sex classification tasks using the MoviNet-A5-Base model with silhouette data type and instances of the WoJ class.

for MoViNet-A0-Base, leads to better results, possibly due to the presence of more instances in training or because some instances are better classified in one of the classes. Finally, it should be noted that except for optical flow obtained with TVL1, an increase in model size improves classification metrics.

Regarding the classification of sex using gait parameters, it is verified whether the results obtained through the parameters acquired with the Optogait system and those estimated from the pose information contain useful patterns for classifying the participants' sex. As previously mentioned, the models used for this are MLP and XGBoost. For the MLP model, binary cross-entropy has been used as the loss function along with a Sigmoid activation function in the neuron of the output layer. The models have been trained for a maximum of $1,000$ epochs, using early stopping during the training process with a patience of 100. On the other hand, for XGBoost, logistic regression is used as the objective with an early stopping of 40 iterations. Please note that the dimensionality of the input data vector is four for the usual gait speed parameters, four for the fast gait speed parameters and three for the circumference parameters, defined in Sec. Methods, Subsec. Data acquisition.

In Table 3, the results of each of the methods are shown. The first rows of both methods show the results using the different variations of the gait parameters. It is interesting to note that in all cases, the use of the parameters estimated from the pose presents better results than those obtained from the parameters measured by the sensors, obtaining the best results with XGBoost using the parameters of the fast gait with an accuracy value of 75.1%. It may be due to the differences in how the various gait parameters are calculated and aggregated into a single value using OptoGait compared to our method. In other words, the descriptors used for the different

Data type	Model	Class	Accuracy	F1 Score
Silhouette	MoViNet A0	WoJ	0.878	0.877
		WJ	0.881	0.881
		Both	0.892	0.892
	MoViNet A5	WoJ	0.918	0.918
		WJ	0.901	0.901
		Both	**0.919**	**0.919**
Segmentation	MoViNet A0	WoJ	0.844	0.844
		WJ	0.854	0.853
		Both	0.846	0.845
	MoViNet A5	WoJ	0.835	0.834
		WJ	0.843	0.842
		Both	0.858	0.858
GMFlow	MoViNet A0	WoJ	0.844	0.844
		WJ	0.836	0.835
		Both	0.855	0.855
	MoViNet A5	WoJ	0.864	0.864
		WJ	0.844	0.843
		Both	0.878	0.878
TVL1	MoViNet A0	WoJ	0.859	0.858
		WJ	0.830	0.829
		Both	0.860	0.860
	MoViNet A5	WoJ	0.846	0.846
		WJ	0.834	0.833
		Both	0.844	0.844

Table 2: Results obtained for the classification of sex using the information extracted from videos for the different combinations of the classes (WoJ, WJ or Both), with the MoviNet-A0-Base and MoviNet-A5-Base models using several frames equal to 40.

machine learning algorithms can better extract helpful information from our representation; that is, they align more closely with the nature of the problem we aim to solve. In addition, the models' training is tested using the waist, hip, and neck circumference parameters. As might be expected, the results obtained show a considerable increase since the measured lengths will be larger in the case of men. However, we are trying to determine whether adding the gait increases the accuracy. It is observed that except for the usual gait parameters for MLP, in the rest of the cases, there is an improvement, reaching, in the best case, an accuracy of 94% using XGBoost with the usual and fast gait parameters. Unlike the previous cases, the best results are obtained from the gait parameters coming from the sensors. However, the results are not very different with respect to the estimated parameters.

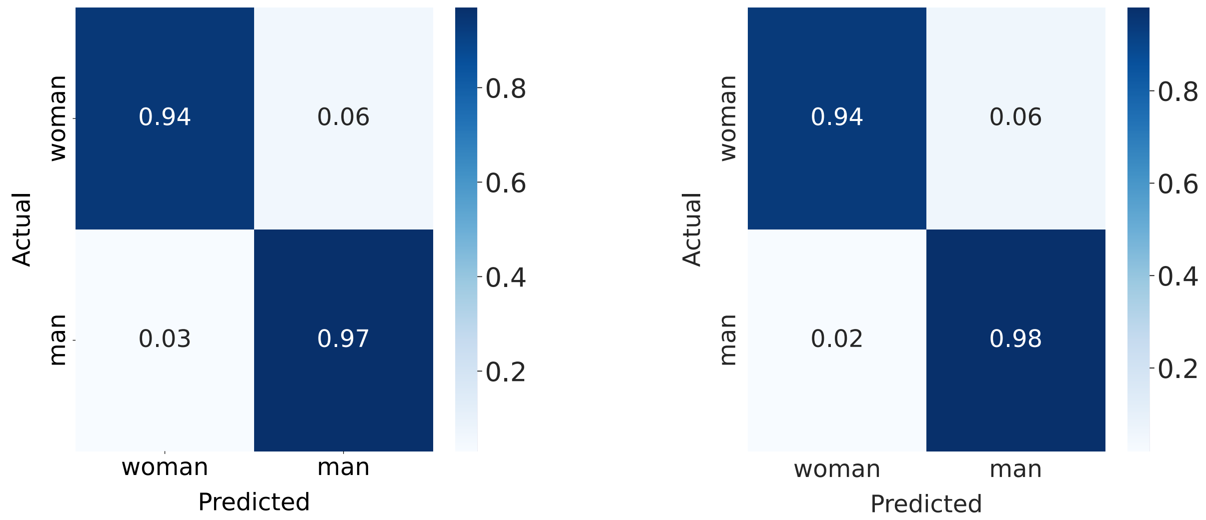
Model	Data Type	Data Source	Accuracy	F1-Score
MLP	Usual Gait parameters	OptoGait	0.512	0.510
		Video	0.714	0.711
	Fast Gait parameters	OptoGait	0.646	0.645
		Video	0.743	0.743
	Usual + Fast Gait parameters	OptoGait	0.668	0.667
		Video	0.749	0.748
	Waist, Hip, Neck circumference	SECA201	0.910	0.909
	Usual Gait + circumferences	OptoGait + SECA 201	0.895	0.895
		Video + SECA201	0.900	0.899
	Fast Gait + circumferences	OptoGait + SECA 201	0.937	0.937
		Video + SECA201	0.912	0.912
	Usual + Fast Gait + circumferences	OptoGait + SECA201	0.937	0.937
		Video + SECA201	0.897	0.897
XGBoost	Usual Gait parameters	OptoGait	0.545	0.525
		Video	0.703	0.702
	Fast Gait parameters	OptoGait	0.638	0.645
		Video	0.751	0.741
	Usual + Fast Gait parameters	OptoGait	0.643	0.649
		Video	0.741	0.737
	Waist, Hip, Neck circumference	SECA201	0.905	0.905
	Usual Gait + circumferences	OptoGait + SECA201	0.927	0.928
		Video + SECA201	0.917	0.916
	Fast Gait + circumferences	OptoGait + SECA201	0.935	0.935
		Video + SECA201	0.937	0.937
	Usual + Fast Gait + circumferences	OptoGait + SECA 201	**0.940**	**0.940**
		Video + SECA201	0.935	0.935

Table 3: Results obtained for the classification of sex using the gait parameters from OptoGait, the parameters estimated from the pose information and circumference measurements, with the MLP and XGBoost models.

Figure 1 display the confusion matrices of the best results using both the silhouette information and that obtained from the gait parameters. It is observed that almost all instances are correctly classified, with slightly more confusion in the 'woman' class.

Figure 2 presents some classification failures of the best models obtained from the silhouette and optical flow. Some of the most notable classification failures using the silhouette model are confusing women with short hair with men and classifying some men wearing sweatshirts with women. The latter case may be due to the model confusing sweatshirt caps with women's long hair. For the case of optical flow, no clear pattern is observed in the misclassification errors.

It could, therefore, be concluded that within the silhouette information, semantic segmenta-



(a) MoViNet-A5-Base on data partition 1 with silhouette information.

(b) XGBoost on data partition 1 with usual and fast gait information, and Waist, Hip, and Neck circumferences.

Figure 1: Confusion matrices for sex classification of the best models obtained (a) using visual information and (b) from gait parameters.

tion, optical flow and gait parameters, information about the sex of the participants is encoded.

Baselines for weight regression.

Similar to previous experiments, this experiment tests whether information about an individual's weight is contained in the various types of data present in the dataset.

Regarding the data corresponding to silhouette, semantic segmentation, and optical flow, to obtain the baselines the MoViNet-A5-Base architecture was used, but modifications were made to the final layers to address a regression task. The backbone is pretrained on the Kinetics 600 dataset. Its output was processed through a 3D Global Average Pooling layer, which then fed into a Multilayer Perceptron (MLP) with a hidden layer of 512 neurons featuring the ReLU activation function, and an output layer comprising a single neuron with a Sigmoid activation function. The backbone was kept frozen throughout the training, updating only the weights of the MLP. The optimiser used was Adam, with a learning rate of 0.0001, and the loss function employed was the Mean Squared Error (MSE). The values to be predicted were normalised within the range of $[0, 1]$. The training was conducted over 100 epochs with Early Stopping set at a patience value of 15, using a batch size of 32. The number of epochs, batch size, and learning rate were selected through a hyperparameter tuning process using cross-validation on validation data. As in the previous experiment, the number of frames has been set to 40.

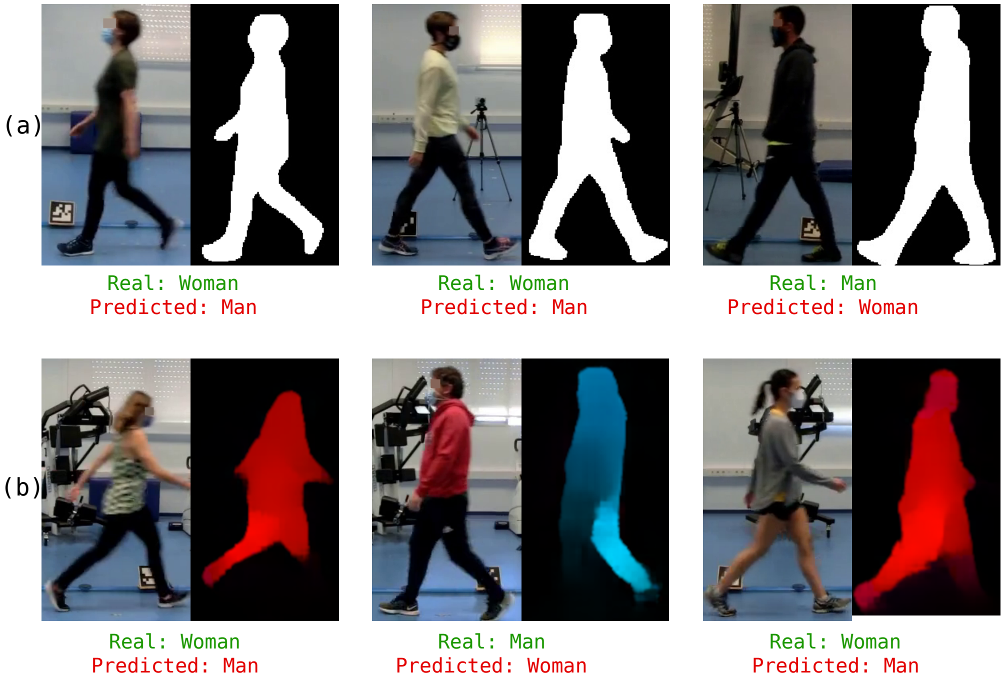


Figure 2: Examples of different sex classification errors for two types of data, silhouette (a) and optic flow (b).

At the top of Table 4, the results obtained for weight estimation using MoViNet-A5-Base are shown, where the best results are obtained from semantic segmentation information with an MAE equal to 6.365 ± 6.207. It is noted that the best results are obtained using semantic segmentation, followed by silhouette representation. To more closely examine where the model errors occur, Figure 3a displays the frequency of relative error percentages for the estimated weight. A relative representation is used because, in the case of weight estimation, making an error in estimating the weight of someone overweight is not as significant as it is for someone thin. It is observed that the maximum errors obtained reach 40% for a small subset of samples. Depending on the level of error one is willing to tolerate, it can be concluded that these types of data representations contain useful information for estimating participants' weight.

On the other hand, it is verified whether the gait parameters, both obtained with OptoGait and those estimated, contain useful information when it comes to estimating the participant's weight. For this purpose, as in the previous case, MLP and XGBoost models are used. The loss function used for training is now the MSE, which is the only change compared to the previous experiment.

In Table 4, shows that the best model obtained is MLP using the parameters of the normal gait estimated from the pose with a value of MAE equal to 10.493 ± 9.757. It is observed that the errors obtained from the gait parameters are higher than those obtained from the visual information extracted from the clips. However, although this is an unexpected result, it is noted that the errors obtained from the gait parameters do not differ excessively either, sometimes falling within the same error intervals. In Figure 3b, shows the relative error for the best model obtained with the gait parameters. The frequencies are lower than in Figure 3a, since in this case there is one instance for each participant, while in the other case, there is one instance for each video in the dataset. However, we observe a behaviour quite similar to the previous case, where the maximum errors obtained reach 40% for a small subset of samples. Therefore, it could be concluded that the gait parameters encode interesting information for estimating the weight of the participants.

Architecture	Data Type	Data Source	MAE (kg) + std
MoviNet A5	Silhouette	Visual	6.506 ± 6.27
	Segmentation	Visual	**6.365 ± 6.207**
	GMFlow	Visual	8.065 ± 7.87
	TVL1	Visual	8.216 ± 7.686
MLP	Usual Gait parameters	OptoGait	12.506 ± 11.065
		Visual	**10.493 ± 9.757**
	Fast Gait parameters	OptoGait	13.022 ± 11.989
		Visual	10.899 ± 10.359
	Usual + Fast Gait parameters	OptoGait	11.869 ± 10.623
		Visual	10.613 ± 9.92
XGBoost	Usual Gait parameters	OptoGait	11.766 ± 10.124
		Visual	10.742 ± 9.871
	Fast Gait parameters	OptoGait	11.680 ± 10.160
		Visual	11.108 ± 10.052
	Usual + Fast Gait parameters	OptoGait	11.729 ± 10.283
		Visual	10.663 ± 9.889

Table 4: Results obtained for weight regression using the information extracted from videos, with MoviNet-A5-Base model, and the gait parameters from OptoGait and the parameters estimated from the pose information, with the MLP and XGBoost models. The best model is highlighted in bold using visual information and using the gait parameters on the other hand.

Baselines for age regression.

In this case, we aim to determine if there is information about a person's age contained in their gait patterns. For this, the same architectures from the previous experiment were used.

Table 5 shows that the best results were obtained using silhouette, with an average error of 9.576 ± 6.662. However, Figure 3c reveals that the percentage of relative errors is very high in many cases, even exceeding 100% in one participant, which highlights the difficulty of the problem being addressed. This could lead to the conclusion that using such data representations, along with other types of characteristics, could be very useful for inferring the age of participants.

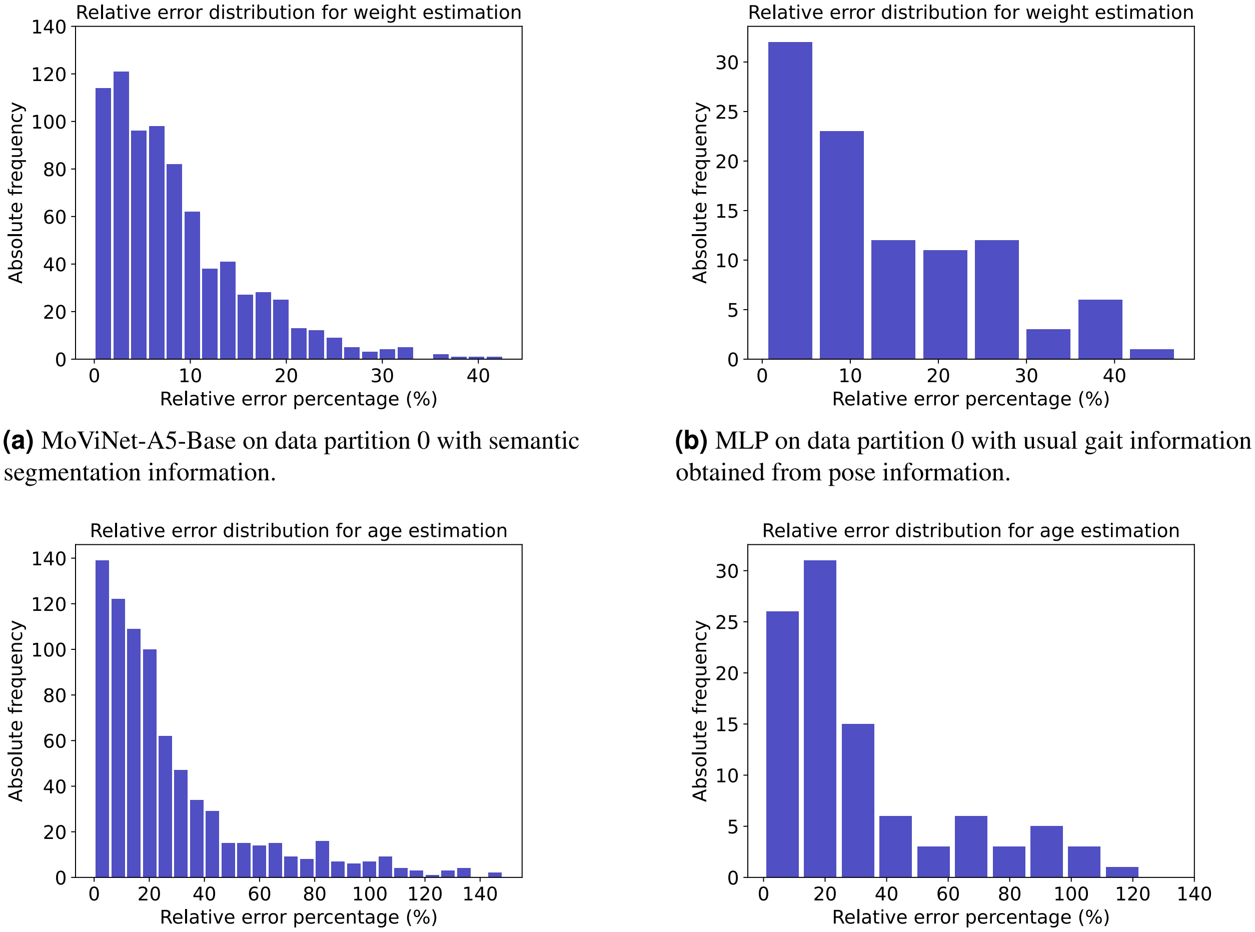
Similar to the previous case, the use of gait parameters leads to an increase in error, being the MAE of the best result obtained 10.409 ± 6.622, which in this instance is less significant than for the estimation of weight. It is again observed that the results obtained from the estimations are better in all cases. In Figure 3d, the distribution of the relative error for the most effective model in data partition two is depicted. Similarly to the weight, the sample size is smaller since it corresponds to the number of participants rather than the number of videos. It is observed that, similar to Figure 3c, the distribution of the relative error is more dispersed compared to the weight estimation, in some cases even exceeding 100% relative error.

Model	**Data Type**	**Data Source**	**MAE (years) + std**
MoviNet A5	Silhouette	Visual	**9.576 ± 6.662**
	Segmentation	Visual	10.066 ± 6.435
	GMFlow	Visual	9.798 ± 6.774
	TVL1	Visual	9.885 ± 6.697
MLP	Usual Gait parameters	OptoGait	11.577 ± 7.115
		Visual	11.09 ± 6.732
	Fast Gait parameters	OptoGait	11.620 ± 6.958
		Visual	11.213 ± 7.676
	Usual + Fast Gait parameters	OptoGait	11.643 ± 7.187
		Visual	10.663 ± 6.744
XGBoost	Usual Gait parameters	OptoGait	11.378 ± 6.540
		Visual	11.134 ± 6.490
	Fast Gait parameters	OptoGait	11.415 ± 6.572
		Visual	10.952 ± 6.689
	Usual + Fast Gait parameters	OptoGait	11.352 ± 6.480
		Visual	**10.409 ± 6.622**

Table 5: Results obtained for age regression using the information extracted from videos, with MoviNet-A5-Base model, the gait parameters from OptoGait and the parameters estimated from the pose information, with the MLP and XGBoost models. The best model is highlighted in bold using visual information and using the gait parameters on the other hand.

References

[1] Kondratyuk, D. et al. MoViNets: Mobile video networks for efficient video recognition. In 2021 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, CVPR 2021,



(a) MoViNet-A5-Base on data partition 0 with semantic segmentation information.

(b) MLP on data partition 0 with usual gait information obtained from pose information.

(c) MoViNet-A5-Base on data partition 2 with silhouette information.

(d) XGBoost on data partition 2 with usual and fast gait information obtained from pose information.

Figure 3: Relative error distribution for weight and age estimation of the best models obtained using visual information and from gait parameters.

16015–16025 (2021).

[2] Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 785–794 (2016).

[3] Head, T., Kumar, M., Nahrstaedt, H., Louppe, G. & Shcherbatyi, I. scikit-optimize/scikit-optimize (2020). URL `https://doi.org/10.5281/zenodo.4014775`.

[4] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C. & Zisserman, A. A short note about kinetics-600. arXiv preprint arXiv:1808.01340 (2018).

[5] Ramachandran, P., Zoph, B. & Le, Q. V. Searching for activation functions. In 6th International Conference on Learning Representations (ICLR 2018) (2018).