Baracca: a Multimodal Dataset for Anthropometric Measurements in Automotive

Stefano Pini, Andrea D'Eusanio, Guido Borghi, Roberto Vezzani, Rita Cucchiara
Department of Engineering "Enzo Ferrari"
University of Modena and Reggio Emilia, Italy
{s.pini, andrea.deusanio, guido.borghi, roberto.vezzani, rita.cucchiara}@unimore.it

Abstract

The recent spread of depth sensors has enabled new methods to automatically estimate anthropometric measurements, in place of manual procedures or expensive 3D scanners. Generally, the use of depth data is limited by the lack of depth-based public datasets containing accurate anthropometric annotations. Therefore, in this paper we propose a new dataset, called Baracca, specifically designed for the automotive context, including in-car and outside views. The dataset is multimodal: it has been acquired with synchronized depth, infrared, thermal and RGB cameras in order to deal with the requirements imposed by the automotive context. In addition, we propose several baselines to test the challenges of the presented dataset and provide considerations for future work.

1. Introduction

The ability to estimate anthropometric measurements – *e.g*. body height, shoulder span, arm length – is a key element in many real world applications and academic research fields, such as soft-biometrics [6], medical health diagnosis [28], person (re)-identification [1], ergonomics [7] and human computer interaction [22].
Usually, accurate anthropometric measurements are collected by qualified personnel (*e.g*. medical staff) relying on time-consuming contact-based measuring methods. Some methods and commercial software that automatically gather anthropometric measurements are available, but they are generally based on high-quality and expensive 3D scanners. In both cases, the measurement accuracy is strongly related to complex acquisition procedures.

Recently, the spread of cheap but accurate active depth sensors, *i.e*. range sensors coupled with an infrared-light emitter, have introduced the possibility to easily and affordably estimate anthropometric measurements. However, a significant issue is represented by the lack of real world and released-for-free datasets containing accurate anthropometric measurements and depth data.

In this paper we present *Baracca*, a new challenging and multimodal dataset collected for the estimation of anthropometric measurements. The dataset consists of more than 9k frames collected by synchronized depth, infrared, thermal and RGB cameras, as shown in Figure 1. We focus on the automotive context and investigate two different acquisition settings: in-car and outside views. An automatic estimation of the anthropometric measurements of the driver (and passengers) – approaching or inside the car – can be used to improve in-cabin ergonomics and human-car interaction (for instance, adjusting the position of seats or rear mirrors). However, the automotive context imposes some requirements [4, 20] for a in-car vision-based system:

- **Non-invasivity**: it is crucial that in-cabin devices do not obstruct the gaze and the movements of the driver. To deal with this requirement, the adoption of a vision-based system is probably the best option [12];
- **Small form factor**: since cameras have to been placed inside the car cockpit, (and usually in specific positions, such as behind the steering wheel or next to the rear-view mirror), a small-sized device is needed;
- **Light invariance**: the vision-based system must be able to work also during the night or during bad weather conditions. In this case, the use of infrared emitters and thermal cameras is a suitable solution;
- **Real-time performance**: the system speed is a crucial element since it has to estimate measurements then quickly provide an output, improving the interaction between the driver and the car [3].

Considering all these elements, we decide to acquire the outside and in-car view sequences of the *Baracca* dataset with multiple sensing devices including, as mentioned above, depth, infrared and thermal sensors in addition to a standard RGB camera. The dataset is publicly released[1].

[1] https://aimagelab.ing.unimore.it/go/baracca

Furthermore, we present several approaches for the anthropometric estimation in order to assess the challenges of the proposed dataset and provide useful baselines for future investigations. In particular, we investigate a geometric-based approach and techniques that belong to the machine learning and the deep learning field.

The rest of the paper is organized as follows: Section 2 presents an overall description of related literature datasets for the task of anthropometric measurement estimation. In Section 3, the proposed dataset, *Baracca*, is detailed. Section 4 presents multiple baselines and reports the experimental results obtained with them. Finally, in Section 5, considerations are drawn.

2. Related Dataset

As mentioned above, in the literature there is a lack of depth-based public datasets acquired for the anthropometric measurement estimation task. At the time of writing, no real-world datasets containing multimodal data are publicly available. To deal with this lack, several methods exploit or generate synthetic datasets that easily recorded and annotated with ground truth measurements.

In [24], three different datasets are introduced but not publicly released. Two datasets are synthetically created, starting from the *MPII Human Shape Model* described in [23] to obtain the 3D model of the human body.
The first proposed dataset contains subjects with the same pose but different body shapes, while in the second one both pose and shape vary. Ground truth anthropometric measurements are obtained using geodesic distances on meshes and body joints; they include body height, shoulder width, leg and foot length, as well as a set of circumferences and thicknesses. Using a virtual depth camera, depth maps are collected through simulation aiming to mimic the projection and the noise of real depth sensors. The real-world dataset include 20 subjects wearing clothes in upright and lie-down poses. The first version of the *Microsoft Kinect* sensor is exploited.

The *CAESAR 3D Anthropometric Database* [25] includes measurements for 2k American and European subjects. It consists in 3D model scans and anthropometric measurements. For each subject a complete 3D models is provided and scanned poses include standing and seated poses. This dataset is available upon payment of a fee.
A variety of full-body 3D scans, captured with an expensive laser scanner, is introduced in [14]. The database contains scans of 59 males and 55 females, which are all fit in a single 3D template model. In [17] a synthetic dataset is introduced, but strongly limited in shape and body variations.
In [29] a small dataset is proposed, in which only 4 subjects are acquired through the first version of the *Microsoft Kinect*. Each subject is standing in the T pose in four different acquisitions: facing the device, in profile, facing away from the camera and halfway between profile and frontal.
In [2] a method to estimate the body height, exploiting the earth gravity, is proposed. In addition, a novel dataset is presented, but it contains only RGB videos of jumping subjects and assuming asymmetric and articulated poses.

There exists some works focused only on specific anthropometric measurements. For instance, [13] proposed a method to estimate the body height, taking into account only the face, starting from the assumption that vertical proportions are constant during the human growth and then they can be exploited to approximate the final measurement. This method relies on an accurate camera calibration procedure. In [21] authors proposed to exploit the knowledge about the pose of the acquisition device – *i.e.* height and pitch angle of the camera with respect to the ground – to regress the height of the acquired body or object.

3. *Baracca* dataset

Baracca is a dataset specifically collected for the anthropometric measurements estimation task on the human body. To the best of our knowledge, this is the first publicly-released dataset that contains depth, infrared, thermal and RGB images, along with manually-collected human body measurements. An overview of the dataset is reported in Figure 1.

3.1. Employed cameras

Considering the requirements imposed by the automotive context, two different cameras have been exploited to collect the data:

- **Pico Zense DCAM710**[2]: this is a depth sensor, based on the *Time-of-Flight* technology, that guarantees the collection of high-quality and low-noise images, especially w.r.t. the *Structured Light* technology [26]. The spatial resolution of the RGB sensor is 1920×1080 pixels, while the infrared/depth sensor has a resolution of 640×480 pixels. The camera is able to acquire valid depth data in the range of 0.2 - 5 meters up to 30 fps. This device is suitable for the automotive context, due to its small form factor ($103 \times 33 \times 22$ mm) and low power consumption (2.5-7.5W). Moreover, the field of view of the infrared/depth sensor ($69°$ horizontal, $51°$ vertical) is suitable for tight spaces.
- **PureThermal 2**[3]: this is a camera board equipped with a *FLIR Lepton* 3.5[4], a low-resolution (160×120 pixels) thermal radiometric sensor which runs at 9 fps. Since the device is radiometric, it is possible to retrieve the

[2] https://www.picozense.com/en/spec.html?spec=710
[3] https://groupgets.com/manufacturers/getlab/products/purethermal-2-flir-lepton-smart-i-o-module
[4] https://groupgets.com/manufacturers/flir/products/lepton-3-5

Figure 1: Overview of the proposed *Baracca* dataset. Rows contain RGB, infrared (IR), depth, and thermal data; columns contain different acquisition points of view (5 indoor views, 3 in-car views).

	Height	**Eye Height**	**Forearm**	**Arm**	**Shoulders**	**Torso**	**Leg**	**Age**	**Weight**	**BMI**
Mean	175.2	164.6	25.73	26.67	42.27	38.63	103.8	26.57	72.03	23.35
Std. dev.	7.100	7.059	1.879	2.134	3.255	2.702	5.536	3.981	12.71	3.222

Table 1: Dataset statistics. Measures of distance, age, weight, and BMI are respectively in cm, years, kg, kg/m^2.

temperature value from the thermal images.
Also in this case, the sensor is suitable for the automotive context: it is self-powered (micro USB, up to 2.5W) and it has a small form factor ($22 \times 30 \times 8$ mm).

3.2. Acquisition procedure

We synchronously collect data from the 2 presented cameras, recording images of 4 different data types, for 30 subjects (26 males, 4 females). Each acquisition contains 5 indoor and 3 in-car sequences (in the left and the right part of Figure 1, respectively) from multiple points of view.
In the outside-view sequences, the subject stands in front of the acquisition devices at different distances. The first two sequences are recorded at 0.6 meters with two different camera viewpoints: top-view and frontal. Then, data are collected frontally at other 3 distances: 1, 1.5 and 2 meters.
In the automotive sequences, cameras are placed on the left A pillar, on the rear-view mirror, and behind the steering wheel. Only the upper body part of the subject – the driver – is here visible.
After the acquisition, many anthropometric measures are collected for each subject, as detailed in the next Section.

3.3. Annotations

The following set of anthropometric measurements is provided for each participant: *height*, *shoulder width*, *forearm and arm length*, *torso width*, *leg length* and *eye height from the ground*. We also include some soft-biometric traits: *age*, *sex* and *weight* of each subject. Some statistics are reported in Table 1.

In addition, we automatically annotate and release the body pose of the subjects for each recorded image. Specifically, we release the the position of 15 skeleton joints in (x, y) image coordinates. Joint prediction is performed using *HRNet* [27, 10], a recent human pose estimation method. The network is trained for 210 epochs on the

COCO dataset [19], which contains RGB images only, with severe data augmentation. Please refer to [27] for additional details. Thanks to the adopted augmentation technique, the network is extremely accurate and able to work in various scenarios. Therefore, we employ it to estimate the body joints of the subjects in each recorded image, obtaining accurate human poses on RGB, IR, and thermal images. In the latter case, images have been normalized and converted to 8-bit images before the pose estimation. Since the infrared and the depth images are aligned, annotations obtained on infrared images are valid also on the depth images.

4. Testing *Baracca* dataset

In this Section, we present some estimation methods that make use of the *Baracca* dataset in order to understand the dataset complexity and to provide useful baselines for future work. The methods are trained to predict the anthropometric measurements reported in Section 3.3. Moreover, we further train a deep model, detailed in the following, to predict the soft-biometric traits reported above (*i.e*. age, weight, BMI). For fair experiments, we split the dataset in official cross-subject train and test splits. 24 subjects are included in the training set while 6 subjects (including 1 female) are included in the test one.

4.1. The Geometric approach

This geometric method estimates the distance between the head of the person and the ground and between the eyes and the ground. It works on the depth data of the outside-view sequences.
The required input is a depth map (and the camera calibration parameters) that is converted to a point cloud. Then, the RANSAC algorithm is used to estimate the plane corresponding to the ground (*i.e*. the plane that fits the elements of the point cloud which belong to the ground). Finally, a trivial point-to-plane distance is calculated to retrieve the height and the eye height of the subject.
This method does not require any training. We report results obtained using the entire point cloud ("Geom. (100%)") and using only 1% of the points ("Geom. (1%)").

4.2. The Machine Learning approaches

The following machine learning methods don't exploit directly the images of the dataset, but use the body skeleton (*i.e*. a set of human joints), calculated in every frame with HRNet [27]. After the skeleton estimation, the following set of distances is calculated over it and used as input for the learning models: head-neck, neck-shoulder, shoulder-elbow, elbow-hand, neck-hip, hip-knee. Only the first 3 distances are used for the in-car view, because the lower body, elbows and hands are often not visible. When possible, these measures are calculated as the mean of left and right side of the body.

We evaluate three machine learning methods: *Linear Regression*, *Random Forests* and *AdaBoost*.
The Linear Regression method simply attempts to fit a linear model to the training data as the least-squares solution.
Random decision trees and forests [5] consist of an ensemble of regression decision trees, which are independently trained as in the bagging technique. When testing, the estimation of each tree is averaged to obtain the final result.
Adaptive Boosting [11, 9] (AdaBoost) consists of multiple weak regressors, sequentially trained weighting the training samples based on the errors of the previous weak regressors. The single predictions are combined with a weighted sum to obtain the final estimation.

4.3. The Deep Learning approach

This deep learning-based method directly estimates the anthropometric measures from the visual appearance only. The deep model is composed of *ResNet-18* [15], without the last fully-connected (fc) layer, pre-trained on *ImageNet* [8], which is used as feature extractor. It is followed by a fc layer with 128 units, batch normalization, ReLU activation and dropout (drop probability $p = 0.5$). Finally, a linear layer with size k regresses the k anthropometric measures. The network is trained optimizing the robust *Huber* loss function [16] through *Adam* [18]. The training is executed for 70 epochs with a batch size of 32 and a learning rate of 0.001, which is reduced by a factor of 10 after 50 and then 65 epochs. The input image size is 128×128.

4.4. Results

The results presented in this section are obtained training the proposed baselines on the training set of *Baracca* and testing them on the test set. We further split the dataset in the "Outside view" split, which contains external sequences (at 1, 1.5 and 2 meters), and in the "In-car view", which contains the in-car sequences.

Baseline results, obtained predicting anthropometric measures from depth, IR, RGB, and thermal data, are respectively reported in Tables 2, 3, 4, and 5. We report the *Mean Absolute Error* (MAE) and the standard deviation calculated between the predicted value and the ground truth in centimeters, aggregated for each subject and then on the whole test set. Considering the depth domain, we report the geometrical approach ("Geom."), which exploits the point clouds; the machine learning approaches, which employ the 3D distances between joints (in camera space); and the deep learning method, which analyzes the depth images. The 3D joints are obtained from the 2D image coordinates using the depth values and the camera calibration parameters.
In the other cases (IR, RGB, and thermal), we report the machine learning approaches, which exploits the 2D distances between joints (in image coordinates), and the deep method, which employ normalized images.

Outside view

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
Geom. (100%)	5.576 ± 4.6	4.393 ± 5.0	-	-	-	-	-	4.985 ± 4.8
Geom. (1%)	5.686 ± 4.9	4.570 ± 5.2	-	-	-	-	-	5.128 ± 5.0
LR	3.853 ± 1.9	1.115 ± 0.3	1.740 ± 0.4	2.151 ± 0.3	4.538 ± 0.5	2.597 ± 1.3	2.317 ± 1.0	2.616 ± 0.8
RandomForest	3.238 ± 2.9	1.187 ± 0.9	1.745 ± 1.1	2.593 ± 1.2	3.912 ± 1.3	3.049 ± 2.3	2.235 ± 1.2	2.566 ± 1.6
Adaboost	3.523 ± 2.3	0.814 ± 0.5	1.382 ± 0.6	2.548 ± 1.1	3.993 ± 1.1	2.310 ± 2.0	2.393 ± 1.1	2.423 ± 1.2
Deep Model	5.724 ± 3.1	5.201 ± 3.0	0.840 ± 0.5	2.014 ± 0.6	2.482 ± 0.7	3.613 ± 1.9	3.040 ± 0.9	3.273 ± 1.5

In-car view

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
LR	3.667 ± 1.5	1.031 ± 0.2	1.937 ± 0.3	2.604 ± 0.3	4.408 ± 0.3	3.474 ± 1.4	2.774 ± 0.8	2.842 ± 0.7
RandomForest	4.185 ± 3.0	1.035 ± 0.8	1.890 ± 1.0	2.686 ± 1.6	4.412 ± 2.1	3.211 ± 2.4	2.494 ± 1.1	2.845 ± 1.7
Adaboost	3.973 ± 1.9	0.939 ± 0.3	1.500 ± 0.3	2.283 ± 0.9	4.627 ± 0.8	3.210 ± 1.4	2.441 ± 0.7	2.710 ± 0.9
Deep Model	7.082 ± 5.9	6.316 ± 5.5	1.016 ± 0.9	2.072 ± 0.9	2.874 ± 1.3	4.734 ± 3.4	3.370 ± 1.4	3.923 ± 2.8

Table 2: Results on the depth domain (MAE ± std (cm)). The ML approaches employ 3D joints in this setting.

Outside view - known distance

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
LR	3.524 ± 2.6	1.020 ± 0.5	1.415 ± 0.6	2.096 ± 0.4	4.108 ± 0.6	2.156 ± 2.2	1.946 ± 1.1	2.324 ± 1.1
RandomForest	3.530 ± 3.0	1.014 ± 0.8	1.973 ± 1.0	2.389 ± 1.1	4.345 ± 1.4	2.649 ± 2.2	2.212 ± 1.2	2.587 ± 1.5
Adaboost	3.998 ± 2.5	0.894 ± 0.4	1.440 ± 0.4	2.317 ± 1.0	4.271 ± 0.8	2.144 ± 1.9	2.147 ± 1.0	2.459 ± 1.1

Outside view

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
LR	5.823 ± 2.1	1.150 ± 0.4	1.629 ± 0.4	2.228 ± 0.3	4.409 ± 0.3	3.412 ± 1.9	2.589 ± 1.4	3.034 ± 1.0
RandomForest	3.916 ± 3.1	1.097 ± 1.1	1.856 ± 0.9	2.737 ± 1.4	4.653 ± 1.3	2.973 ± 2.7	2.250 ± 1.3	2.783 ± 1.7
Adaboost	4.542 ± 1.8	1.011 ± 0.3	1.253 ± 0.4	2.328 ± 0.6	4.480 ± 0.9	3.117 ± 1.6	2.253 ± 1.0	2.712 ± 0.9
Deep Model	6.641 ± 1.9	5.914 ± 1.8	1.109 ± 0.3	1.965 ± 0.3	2.579 ± 0.4	4.113 ± 1.1	3.011 ± 0.4	3.619 ± 0.9

In-car view

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
LR	4.975 ± 1.0	0.964 ± 0.3	1.837 ± 0.4	2.527 ± 0.5	4.412 ± 0.6	3.509 ± 1.1	2.885 ± 0.8	3.016 ± 0.6
RandomForest	6.515 ± 3.5	1.151 ± 1.1	2.056 ± 1.3	2.493 ± 1.7	4.396 ± 1.8	3.822 ± 2.5	2.809 ± 1.4	3.320 ± 1.9
Adaboost	4.924 ± 1.3	1.018 ± 0.2	1.395 ± 0.4	2.408 ± 0.5	4.395 ± 0.6	4.206 ± 2.0	3.015 ± 0.7	3.052 ± 0.8
Deep Model	6.555 ± 3.6	6.488 ± 3.4	1.130 ± 0.6	2.034 ± 0.7	1.895 ± 0.9	3.952 ± 2.2	3.022 ± 1.0	3.582 ± 1.8

Table 3: Results on the IR domain (MAE ± std (cm)). The ML approaches employ 3D joints in the "known distance" setting, 2D joints otherwise.

Moreover, in the IR and RGB case, we further report results obtained using the 3D distances between joints in the "Outside view". We exploit the known distance (1, 1.5, 2 meters) as depth approximation and the camera calibration parameters to convert the 2D joints (in image coordinates) to the 3D ones (in camera space).

In addition, Table 6 contains the results obtained by the deep model trained for the estimation of soft-biometric traits. Then, Table 7 presents the inference time of the proposed approaches.

5. Discussion

In this paper we have presented *Baracca*, a multimodal dataset for the estimation of anthropometric measures. Results in Tables 2, 3, 4, 5 show that these measurements can be successfully estimated using any of the data types included in the dataset (*i.e.* depth, infrared, RGB and thermal). In every case, the machine learning approaches, which exploit the accurate joint prediction of HRNet [27], are the best-performing solution.

Considering the IR and RGB domain (Tables 3 and 4), the use of approximate 3D joints further improves the accuracy of these methods, confirming that 3D data, independent from the camera intrinsics, are the most suitable data for the anthropometric estimation. However, with this kind of sensors the 2D to 3D conversion is possible only if the distance between the subject and the camera is known and if the subject joints can all be considered at the same distance.

In view of this, the most adequate sensors for anthropometric estimation are the depth ones, which naturally gather the 3D information of the scene (Table 2). Using this type of data, even a simple geometrical approach can be employed, in addition to the other ones, obtaining acceptable, but less accurate results. It is worth to note that this approach still obtain low MAE even with extremely-small point clouds (1% of the original one, consisting in just 1k-2k points), al-

Outside view - known distance

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
LR	3.633 ± 2.0	1.089 ± 0.5	1.647 ± 0.3	1.981 ± 0.4	4.624 ± 0.6	2.002 ± 1.5	1.587 ± 1.1	2.366 ± 0.9
RandomForest	3.844 ± 2.2	1.147 ± 0.9	1.888 ± 0.8	1.982 ± 1.0	4.740 ± 1.2	2.734 ± 1.9	1.882 ± 1.1	2.602 ± 1.3
Adaboost	2.877 ± 1.8	0.955 ± 0.5	1.455 ± 0.4	1.995 ± 0.7	4.610 ± 0.5	2.442 ± 1.8	1.931 ± 0.7	2.324 ± 0.9

Outside view

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
LR	5.260 ± 1.8	1.007 ± 0.4	1.866 ± 0.4	2.201 ± 0.5	4.859 ± 0.5	3.409 ± 1.4	2.510 ± 1.5	3.016 ± 0.9
RandomForest	4.321 ± 2.8	1.082 ± 1.0	1.784 ± 0.9	1.946 ± 1.1	4.801 ± 1.2	2.891 ± 2.6	2.189 ± 1.2	2.716 ± 1.5
Adaboost	4.771 ± 1.4	1.004 ± 0.2	1.280 ± 0.3	2.192 ± 0.7	4.716 ± 0.6	3.033 ± 1.2	2.235 ± 0.8	2.747 ± 0.7
Deep Model	10.124 ± 3.4	9.373 ± 3.2	1.564 ± 0.5	1.898 ± 0.5	2.355 ± 0.7	6.392 ± 2.0	3.348 ± 0.9	5.008 ± 1.6

In-car view

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
LR	5.224 ± 1.3	0.955 ± 0.3	1.775 ± 0.3	2.477 ± 0.4	4.365 ± 0.5	3.770 ± 1.1	2.811 ± 0.8	3.054 ± 0.7
RandomForest	6.481 ± 3.7	1.307 ± 1.0	2.240 ± 1.2	2.724 ± 1.7	4.278 ± 1.3	4.937 ± 3.1	3.466 ± 1.6	3.633 ± 1.9
Adaboost	6.014 ± 1.5	0.912 ± 0.3	1.541 ± 0.4	2.279 ± 0.6	4.279 ± 0.3	4.378 ± 1.3	3.093 ± 1.2	3.214 ± 0.8
Deep Model	7.898 ± 3.5	7.810 ± 3.3	1.520 ± 0.6	2.115 ± 0.6	2.314 ± 0.7	4.973 ± 1.9	2.776 ± 0.8	4.201 ± 1.6

Table 4: Results on the RGB domain (MAE ± std (cm)). The ML approaches employ 3D joints in the "known distance" setting, 2D joints otherwise.

Outside view

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
LR	4.877 ± 1.4	1.182 ± 0.4	1.496 ± 0.5	2.316 ± 0.5	4.117 ± 0.7	3.347 ± 1.2	2.607 ± 0.9	2.849 ± 0.8
RandomForest	5.278 ± 3.5	1.351 ± 1.0	1.568 ± 0.8	2.268 ± 1.3	3.861 ± 1.1	3.767 ± 2.8	2.294 ± 1.3	2.912 ± 1.7
Adaboost	5.064 ± 1.9	1.131 ± 0.4	1.349 ± 0.4	2.250 ± 0.7	4.291 ± 0.5	3.203 ± 1.8	2.442 ± 1.0	2.819 ± 1.0
Deep Model	5.267 ± 3.1	4.939 ± 2.9	0.955 ± 0.4	2.220 ± 0.5	2.458 ± 0.7	3.659 ± 1.8	2.930 ± 0.8	3.204 ± 1.4

In-car view

Method	Height	Eye Height	Forearm	Arm	Torso	Leg	Shoulders	Average
LR	4.823 ± 1.2	1.087 ± 0.2	1.611 ± 0.1	2.251 ± 0.3	4.409 ± 0.5	3.112 ± 1.0	2.443 ± 0.5	2.819 ± 0.6
RandomForest	5.038 ± 3.7	1.402 ± 1.0	1.826 ± 0.9	2.233 ± 1.2	4.678 ± 1.3	3.365 ± 2.9	3.035 ± 1.6	3.082 ± 1.8
Adaboost	4.856 ± 2.0	1.172 ± 0.2	1.792 ± 0.5	2.295 ± 0.6	4.587 ± 0.5	3.507 ± 1.3	2.805 ± 1.1	3.002 ± 0.9
Deep Model	6.632 ± 2.7	6.320 ± 2.7	0.945 ± 0.4	2.317 ± 0.4	2.441 ± 0.7	4.542 ± 1.5	3.479 ± 0.7	3.811 ± 1.3

Table 5: Results on the thermal domain (MAE ± std (cm)). The ML approaches employ 2D joints.

Outside view

Domain	Age	Weight	BMI
Depth	3.863 ± 0.8	10.749 ± 5.0	3.247 ± 1.3
IR	3.824 ± 0.6	5.689 ± 3.4	2.278 ± 0.9
RGB	3.530 ± 0.6	16.537 ± 4.8	4.098 ± 1.3
Thermal	4.040 ± 0.8	9.926 ± 3.8	2.386 ± 1.0

In-Car view

Domain	Age	Weight	BMI
Depth	3.819 ± 0.8	9.561 ± 6.4	2.603 ± 1.5
IR	4.914 ± 1.6	7.410 ± 3.2	2.235 ± 0.8
RGB	4.135 ± 1.0	11.959 ± 5.4	2.992 ± 1.4
Thermal	3.550 ± 0.9	11.012 ± 5.7	2.620 ± 1.5

Table 6: Age, weight and BMI estimated by the Deep Model (Sec. 4.3) using different domains (MAE ± std).

Method	Inference time (ms)	
	CPU	GPU
Geom. (100%)	741.9 ± 138.3	-
Geom. (1%)	66.81 ± 1.992	-
HRNet	591.8 ± 134.2	61.81 ± 24.44
+ LR	0.047 ± 0.006	-
+ RandomForest	0.540 ± 0.013	-
+ Adaboost	1.527 ± 0.693	-
Deep Model	23.07 ± 0.430	4.619 ± 0.289

Table 7: Inference time of the tested approaches (ms ± std).

lowing the use cheap low-resolution depth sensors.

Regarding the inference time, as it can be seen in Table 7, the ML approaches are extremely fast, but require the subject body joints (calculated, for instance, with HRNet [27]) increasing the overall inference time. Therefore, the deep method is the fastest approach, regardless of running it on CPU or GPU. Moreover, this method can estimate additional soft-biometric traits with a relatively-low average error from any data type, as shown in Table 6.

Future work include the development of multimodal and point cloud-based algorithms for anthropometric measurements. In addition, the thermal data could be used for the estimation of the thermal comfort of the car passengers.

References

[1] V. O. Andersson and R. M. Araujo. Person identification using anthropometric and gait data from kinect sensor. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 1

[2] D. Bieler, S. Gunel, P. Fua, and H. Rhodin. Gravity as a reference for estimating a person's height from video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8569–8577, 2019. 2

[3] G. Borghi, E. Frigieri, R. Vezzani, and R. Cucchiara. Hands on the wheel: a dataset for driver hand detection and tracking. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 564–570. IEEE, 2018. 1

[4] G. Borghi, R. Gasparini, R. Vezzani, and R. Cucchiara. Embedded recurrent network for head pose estimation in car. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1503–1508. IEEE, 2017. 1

[5] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. 4

[6] A. Dantcheva, C. Velardo, A. D'angelo, and J.-L. Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, 2011. 1

[7] B. Das and A. K. Sengupta. Industrial workstation design: a systematic ergonomics approach. *Applied ergonomics*, 27(3):157–163, 1996. 1

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 4

[9] H. Drucker. Improving regressors using boosting techniques. In *ICML*, volume 97, pages 107–115, 1997. 4

[10] A. D'Eusanio, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara. Manual annotations on depth maps for human pose estimation. In *International Conference on Image Analysis and Processing*, pages 233–244. Springer, 2019. 3

[11] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995. 4

[12] E. Frigieri, G. Borghi, R. Vezzani, and R. Cucchiara. Fast and accurate facial landmark localization in depth images for in-car applications. In *International Conference on Image Analysis and Processing*, pages 539–549. Springer, 2017. 1

[13] Y.-P. Guan et al. Unsupervised human height estimation from a single image. *Journal of Biomedical Science and Engineering*, 2(06):425, 2009. 2

[14] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Computer graphics forum*. Wiley Online Library, 2009. 2

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[16] P. J. Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992. 4

[17] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. Moviereshape: Tracking and reshaping of humans in videos. *ACM Transactions on Graphics (TOG)*, 29(6):1–10, 2010. 2

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[20] F. Manganaro, S. Pini, G. Borghi, R. Vezzani, and R. Cucchiara. Hand gestures for the human-car interaction: the briareo dataset. In *International Conference on Image Analysis and Processing (ICIAP)*, pages 560–571. Springer, 2019. 1

[21] M. Momeni-k, S. C. Diamantas, F. Ruggiero, and B. Siciliano. Height estimation from a single camera view. In *VISAPP (1)*, pages 358–364, 2012. 2

[22] A. Phinyomark, F. Quaine, and Y. Laurillau. The relationship between anthropometric variables and features of electromyography signal for human–computer interface. In *Applications, Challenges, and Advancements in Electromyography Signal Processing*, pages 321–353. IGI Global, 2014. 1

[23] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67:276–286, 2017. 2

[24] T. Probst, A. Fossati, M. Salzmann, and L. Van Gool. Efficient model-free anthropometry from depth data. In *2017 International Conference on 3D Vision (3DV)*, pages 486–495. IEEE, 2017. 2

[25] K. M. Robinette, H. Daanen, and E. Paquet. The caesar project: a 3-d surface anthropometry survey. In *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No. PR00062)*, pages 380–386. IEEE, 1999. 2

[26] H. Sarbolandi, D. Lefloch, and A. Kolb. Kinect range sensing: Structured-light versus time-of-flight kinect. *Computer vision and image understanding*, 139:1–20, 2015. 2

[27] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 3, 4, 5, 6

[28] N. Utkualp and I. Ercan. Anthropometric measurements usage in medical sciences. *BioMed research international*, 2015, 2015. 1

[29] A. Weiss, D. Hirshberg, and M. J. Black. Home 3d body scans from noisy image and range data. In *2011 International Conference on Computer Vision*, pages 1951–1958. IEEE, 2011. 2