POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Real-time evaluation of the human body's inertia tensor by using a stereo-depth camera and a deep-learning-based algorithm

Laurea Magistrale in Mechanical Engineering - Ingegneria Meccanica

Author: Laura Viprati

Advisor: Prof. Hermes Giberti

Co-advisor: Marco Carnevale, Nicola Giulietti

Academic year: 2023-2024

1. Introduction

Researches on the dynamic contribution of the human body during motion have garnered significant interest over the years. Having a real-time method for estimating the inertia tensor of the human body enables the calculation of mutually exchanged forces during human-robot interaction [1] and it promotes the development of "human-in-the-loop" dynamic simulators [2], which are characterized by the non negligible influence of human motion on the dynamic of the action that is being simulated. With the present thesis it is being addressed the need for a method capable of supplying real-time updates of inertial quantities, placing particular emphasis on its robustness and versatility in terms of adaptability to various human sizes and proportions and physical activities that are being performed. It is developed a geometric model of the human body, which consists in the discretization of each body part into simple three-dimensional geometrical shapes, using a depth camera and a deep-learning-based algorithm to real-time estimate the individual's pose and, consequently, his inertial properties. Several algorithms are employed with the purpose of identifying those that simultaneously provide accurate estimates and approximate real-time conditions effectively, taking into consideration also different acquisition frequencies to highlight the main advantages or drawbacks associated with a decrease in the time interval among consecutive updates.

2. Materials and methods

2.1. Experimental setup

The image acquisition system utilized for the identification of the pose of the human body consists of the stereo-depth camera Intel RealSense D415, which is equipped with an infrared projector and two sensors for the evaluation of the three-dimensional positioning of each pixel. The image made available by the camera is then processed by a deep-learning-based algorithm for the identification of key body landmarks. For the purpose of the present thesis various algorithms have been considered, differing in the number of detectable points and their precise location on the body. MediaPipe Pose is a pre-trained pose estimation model which provides the positioning of 33 three-dimensional key body points, which correspond to various body joints or facial features (Figure 1). TensorFlow and YOLOv8, instead,

represent two sets of models which estimate the position of 17 body landmarks that are coincident with those identified by MediaPipe excluding hands, feet, mouth and inner and outer corner of the eyes.

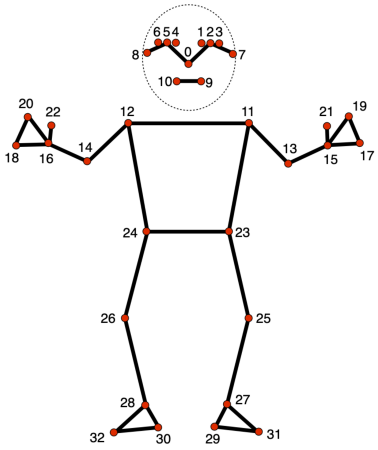

Figure 1: 33 three-dimensional body landmarks identified by MediaPipe.

For a correct capturing of the body pose the camera is placed in front of the individual at about one meter above the ground. It is fundamental to ensure that in the captured image there are no other individuals aside from the one undergoing the experiment and that the camera is placed sufficiently distant from the participant, so that he can move freely without his limbs going beyond the camera's field of view.

2.2. Definition of the geometric model

Body parts are discretized into simple three-dimensional shapes starting from the points' positions identified by the algorithm for the definition of the geometric model. Given the necessity to perform quick calculations to ensure a real-time update of the resulting inertial quantities, this model rely on some simplifying assumptions: each body segment has isotropic density, mass division among body segments and segmental volumes are constant and the body is assumed to be symmetric with respect to the sagittal plane. Starting from the 33 points' positions made available by MediaPipe Pose, the human body is discretized into 14 segments (shown in Figure 2), whose shapes are the same as those adopted in [3]: head and neck are discretized as an ellipsoid, the torso as a cylinder with elliptical cross section, upper arms, lower arms, upper legs and lower legs as frustums of cone and, finally, hands as spheres.

Figure 2: Reproduction of the 14-segmental geometrical model of the human body in CAD

Considering relevant anthropometric studies concerning human body proportions, the dimension of each body part has been set as follows. The semi-minor axis of the ellipsoid is set as the ratio between the distance among the ears and 1.5, while the semi-major one is defined as the product between the distance among eyes and mouth and 2.2. The length of the torso, as well as that of the limbs, is defined directly starting from the points identified by the algorithm. The minor semi-axis of the torso, instead, is equal to 60% of the major one, whose length corresponds to half the distance between the shoulders. The dimension of the bases of the limbs, which are all given the shape of frustums of cone, is derived starting from their heights: the ratio between the height of the segment and the major and minor base circumferences is set respectively equal to 1.1 and 1.2 for the upper arms, 1.2 and 1.7 for the lower arms, 0.85 and 1.1 for the upper legs, 1.1 and 1.55 for the lower legs, 1.55 and 0.8 for the feet. The only remaining body parts are the hands, which are given the shape of spheres whose rays are equal to the ratio between the distance from wrist to thumb and 1.5.
If, instead of using MediaPipe, models based on the 17 landmarks identified by TensorFlow and YOLOv8 were employed for determining the size of each discretized body part, a distinct geometric model would be obtained. The human body

in this case is divided into 10 rather than 14 segments, which are the same as those characterizing the 33-point model excluding hands and feet. For what concerns the size of each part, calculations remain the same with the only exception of the head: since the points' positions estimated do not include the mouth, in this case the minor radius of the ellipsoid is set equal to 58% of the major one.
Every time a new image made available by the camera is processed by one of these algorithms, since they have not been designed with the purpose of keeping constant segments' lengths or proportions, the dimension of each discretized body part may vary. To overcome this issue, it has been chosen to average the length of each body segment considering values obtained over multiple acquisitions.

2.3. Segmental volumes and masses

Segmental volumes can be calculated considering well-known formulas associated with each one of the geometrical shapes adopted for the definition of the geometrical model. For what concerns the mass of each body segment, instead, it is set as a percentage of the total mass of the body using values defined by Dumas in [4], which are listed in Table 1.

	Female	Male
Head and neck	6.7 %	6.7 %
Torso	45.1 %	47.5 %
Upper arm	2.2 %	2.4 %
Lower arm	1.3 %	1.7 %
Hand	0.5 %	0.6 %
Upper leg	14.6 %	12.3 %
Lower leg	4.5 %	4.8 %
Foot	1 %	1.1 %

Table 1: Body segments' mass percentages for the 33-point model.

These values are valid for the 33-point model, while for the 17-point ones it is necessary to re-scale the percentages associating hands' and feet's contributions respectively to the lower part of the arms and the lower part of the legs.

2.4. Evaluation of the position of the center of mass

To evaluate the position of the center of mass (COM) of the entire body, it is necessary at first to define individually the position of the COM of each body part along its axis of symmetry. Regarding the sphere, the ellipsoid and the elliptical cylinder, given the hypothesis of isotropic density, their COMs are located exactly at half of their length, while for the frustums of cone it is utilized the following formula:

$$COM = \frac{h(1/4R^2 + 1/2Rr + 3/4r^2)}{R^2 + Rr + r^2},$$

where h represents the height, R the radius of the major base and r the radius of the minor base. In this way it is obtained the distance of the COM from the bigger base of each frustum. These positions are then re-expressed with respect to a global reference frame, which has been chosen as the one having the origin of x and z axes in correspondence of the mid-point between the hips and the origin of the y axis coincident with the lower among the points identified by the algorithm. The y axis is vertical and perpendicular to the ground, the x axis is directed toward the left of the body and the z axis outward from the body. To prevent the global reference frame's origin from varying with time, it is considered fixed and coincident with the one evaluated for the initial pose of the individual. The position of the COM of the entire body is finally obtained through the following weighted mean:

$$X_{COM} = \frac{\sum_{i=0}^{n} m_i x_i}{M},$$
$$Y_{COM} = \frac{\sum_{i=0}^{n} m_i y_i}{M},$$
$$Z_{COM} = \frac{\sum_{i=0}^{n} m_i z_i}{M}.$$

M is the total body mass, m_i is the mass of the i^{th} segment, x_i, y_i and z_i are respectively the positions along x, y and z of the COM of the i^{th} segment expressed in the global reference frame.

2.5. Evaluation of the inertia tensor

The central objective of the present thesis is the estimation of the inertia tensor of the human body. This process initiates with the computation of the matrix associated with each distinct body segment and is followed by the application of adequate transformations, which include

rotation and translation, with the purpose of aligning the local reference frames assigned to each body part with a shared Common Reference Frame (CRF). The adopted CRF has axes' orientation coincident with that of the global reference frame define previously, but with the origin centered in the COM of the entire body. The rotation of each segmental inertia tensor (I_{local}) is carried out performing the following multiplication:

$$I_{rotated} = DCM * I_{local} * DCM',$$

where DCM represents the direction cosine matrix, which indicates the relative orientation among each body segment's local reference frame and the CRF. The second transformation executed consists in the translation of the rotated inertia matrix into the origin of the CRF through the following operation:

$$I_{translated} = I_{rotated} + I_{translation},$$

where

$$I_{translation} = m \begin{bmatrix} r_y^2 + r_z^2 & -r_x r_y & -r_x r_z \\ -r_y r_x & r_x^2 + r_z^2 & -r_y r_z \\ -r_z r_x & -r_z r_y & r_x^2 + r_y^2 \end{bmatrix}.$$

Vector $\mathbf{r} = [r_x, r_y, r_z]$ contains the distance among the origin of the local and the common reference frames for each body segment.
The overall human body's inertia tensor is finally obtained summing up all the segmental contributions. Since the order followed in the execution of the transformations does not affect the resulting values, rotation and translation operations can be interchanged.

3. Results

3.1. 33-point model

The test has been conducted on a female individual with height and mass respectively equal to 1.58 m and 55 kg. The results in terms of position of the segmental centers of mass and segmental inertia tensors obtained adopting the 33-point model are shown in Table 2. To validate these values it is made a comparative analysis considering the corresponding quantities obtainable applying the regression-based method developed by Zatsiorsky and then re-adjusted by De Leva in [5]. Despite the unavoidable differences associated with the fact that, differently from the model developed in this thesis, De Leva evaluated human body's inertial properties starting from multiple anthropometric measurements of various human subjects and adopting gamma-ray scanning techniques to better discretize the dimension of each body part, the results show a good agreement.
Furthermore, in order to validate the mathematical passages presented in the preceding section, the geometrical model has been replicated using the CAD software Inventor. The resulting quantities showed a great correspondence, with a maximum difference equal to 0.1% for the positioning of the COM and 1.4% for the inertia tensor of the entire body. This small divergence is not caused by erroneous mathematical steps, but it is probably related to numerical errors or to the not perfect positioning of the model in CAD.

	Present study				De Leva's study			
	COM	**Ixx**	**Iyy**	**Izz**	**COM**	**Ixx**	**Iyy**	**Izz**
Head and neck	50%	17.85	9.62	17.85	58.94%	18.98	14.89	16.04
Torso	50%	594.81	215.63	696.29	41.51%	753.82	191.81	836.00
Upper arm	48.55%	8.51	0.99	8.51	57.54%	7.18	2.32	8.20
Lower arm	42.32%	3.84	0.38	3.84	45.59%	3.50	0.47	3.61
Hand	50%	0.14	0.14	0.14	74.74%	0.39	0.21	0.53
Upper leg	46.05%	113.57	19.06	113.57	36.12%	146.26	28.97	150.30
Lower leg	41.64%	24.63	3.12	24.63	44.16%	35.25	4.28	36.31
Foot	46.31%	0.96	0.29	0.96	40.14%	2.80	0.71	3.31

Table 2: Comparison of segmental centers of mass and principal moments of inertia between the current study and De Leva's research. The COMs are expressed as percentages of each segment's length and the unit of measurement for the moments of inertia is $kg\ m^2 \cdot 10^3$.

3.2. 17-point models

The 17-point models analyzed, together with those identifiable through TensorFlow's and YOLOv8's algorithms, include also the one obtainable reducing the number of points identified by MediaPipe from 33 to 17. The same video showing the motion of a single individual has been processed by both the original and the reduced MediaPipe's models. Considering the resulting inertial quantities presented in Figure 3, it is possible to calculate mean and the standard deviation of the difference between the values obtained for the two models at each time instant. For the positioning of the COM, the mean of the difference along x, y and z is respectively equal to 0.12 cm, 0.73 cm and 0.14 cm, while the standard deviation is equal to 0.09 cm, 0.18 cm and 0.12 cm. For the inertia tensor the mean of the difference about x, y and z axes is respectively equal to 0.67, 0.10 and 0.68 $kg\ m^2$, while the standard deviation is equal to 0.25, 0.07 and 0.16 $kg\ m^2$. These values are representative of the fact that, despite a small difference associated with the simplification of the geometrical model exists, it is not excessively influencing. Besides, considering the precious advantages offered by the adoption of a 17-point model, which include the reduction of computational times and the mitigation of errors associated with hands' and feet's positioning, it could be considered a good choice to opt for this model rather than for the original 33-point one.
Another comparison is done between the resulting inertial quantities obtained applying TensorFlow's and YOLOv8's algorithms and those estimated using MediaPipe's reduced model. The resulting differences are always in the order of cm or even mm for the positioning of the COM, while for the inertia tensor they are in the order of tenths and hundredths of $kg\ m^2$. A notable observation is that YOLOv8 tends to underestimate the moments of inertia compared to MediaPipe, a behavior not observed for TensorFlow's models. The comparative differences can be reduced increasing the acquisition frequency: switching from 30 fps to 90 fps, in fact, despite the lower camera resolution, a greater amount of information becomes available, enabling the system to respond more promptly to a change in body pose and, consequently, in inertial properties.

4. Conclusions

In conclusion the geometrical model developed in this thesis proves to be successful, effectively enabling real-time estimations of the human body's inertial properties through the exploitation of a depth camera and a deep-learning-based algorithm. The association of each body pose with the respective center of mass positioning and inertia tensor is achieved with an accuracy on the order of mm and tenths of $kg\ m^2$. While these values are influenced to some extent by the camera's resolution, the primary limiting factor for a more precise estimation derives from the significant simplification of the human body caused by the adoption of a geometric model.
Concerning the choice of the algorithm to be adopted, it has been concluded that 17-point models, despite losing information related to the positioning of body extremities, are advantageous as they allow for quicker estimation updates, involving smaller computational times,

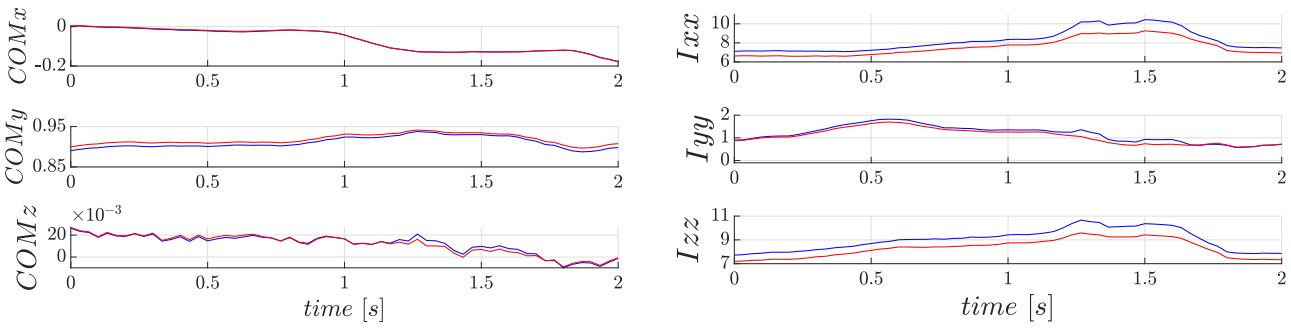

Figure 3: Comparison between the position of the COM and the inertia tensor estimated in real-time using the original 33-point model (blue line) and the reduced MediaPipe's one (red line). The COM is expressed in m, while the inertia tensor in $kg\ m^2$.

and consequently for a better approximation to real-time conditions. Increasing the acquisition frequency from 30 fps to 90 fps, in fact, it is possible to obtain a more faithful estimation, providing greater responsiveness in detecting even smaller and sudden movements of the body. An additional advantage of employing 17-point models rather than 33-point ones lies in the omission of points which are generally associated with larger positioning errors. Hands and feet, in fact, are more likely to fall outside the field of view of the camera, in particular in cases where the individual is performing broad movements. Accordingly, excluding the analysis of these body segments proves beneficial in preventing the identification of approximate or excessively inaccurate models.
Potential avenues for future researches could involve a discretization of the human body that more thoroughly considers the differences among male and female figures, allowing for a more accurate sizing of each body part. Furthermore, a more detailed analysis of the multiple 17-point models available could effectively enable the individuation of the best-performing model, capable of yielding closer-to-reality inertial values.

References

[1] Claudia Latella. Human whole-body dynamics estimation for enhancing physical human-robot interaction. *arXiv preprint arXiv:1912.01136*, 2019.

[2] Katherine Driggs-Campbell, Guillaume Bellegarda, Victor Shia, S Shankar Sastry, and Ruzena Bajcsy. Experimental design for human-in-the-loop driving simulations. *arXiv preprint arXiv:1401.5039*, 2014.

[3] Manisha Jagadale, KN Agrawal, CR Mehta, RR Potdar, and Nandni Thakur. Estimation and validation of body segment parameters using 3d geometric model of human body for female workers of central india. *Agricultural Research*, 11(4):768–780, 2022.

[4] Raphaël Dumas, Laurence Cheze, and J-P Verriest. Adjustments to mcconville et al. and young et al. body segment inertial parameters. *Journal of biomechanics*, 40(3):543–553, 2007.

[5] Paolo De Leva. Adjustments to zatsiorsky-seluyanov's segment inertia parameters. *Journal of biomechanics*, 29(9):1223–1230, 1996.