Careful with That! Robot Observing Human Handlings to Estimate Object Properties

Linda Lastrico[*1,2], Alessandro Carfì[1], Alessia Vignolo[3], Alessandra Sciutti[3], Fulvio Mastrogiovanni[1] and Francesco Rea[2]

[1] Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi (DIBRIS), Università degli Studi di Genova, Genova, Italy
[2] Robotics, Brain and Cognitive Science Department (RBCS), Italian Institute of Technology, Genova, Italy
linda.lastrico@iit.it
[3] Cognitive Architecture for Collaborative Technologies Unit (CONTACT), Italian Institute of Technology, Genova, Italy

Abstract. Humans are very effective at interpreting subtle properties of the partner's movement and use this skill to promote smooth interactions. Therefore, robotic platforms that support human partners in daily activities should acquire similar abilities. In this work we focused on the features of human motor actions that communicate insights on the weight of the object and the carefulness required in its manipulation. Our final goal is to enable a robot to autonomously infer the degree of care required in object handling and to discriminate whether it is light or heavy, just by observing a human manipulation. This preliminary study represents a promising step towards the implementation of those abilities on a robot observing the scene with its camera. Indeed, we succeeded in demonstrating that it is possible to reliably deduct if the human operator is careful when handling a object, through machine learning algorithms relying on the stream of visual acquisition from either a robot camera or from a motion capture system. On the other hand, we observed that the same approach is inadequate to discriminate between light and heavy objects.

Keywords: Biological motion kinematics · Human motion understanding · Natural communication · Deep learning · Human-robot interaction

1 Introduction and background

In the context of human-robot interaction, a great effort is directed towards the development of the robot ability to understand implicit signals and subtle cues that naturally characterize human movements. This comes to have critical importance in situations where robots are used in unconstrained environments, for instance in manufacturing, helping human operators to lift loads or assisting elderly. In typical human-human interaction a considerable amount of information is exchanged through non-verbal signals, such as the attitude of an action, its tempo, the direction of the gaze and the body posture. It has been proved that

people are able to correctly estimate the weight of an object, simply observing another person lifting it [15]. Recent research confirmed that the same information could be transmitted by a humanoid robot controlling the vertical velocity of its lifting movements [8]. Humans manage to easily manipulate objects they have never used before: at first, by inferring their properties such as the weight, the stiffness and the dimensions also from the observation of others manipulating them; at a later time, using tactile and force feedback to improve the estimation. Replicating this behaviour in robotic systems is challenging. However, preliminary results have been achieved in estimating objects physical properties, relying on inference-based vision approaches [13].
The interaction with humanoid robots is particularly critical: driven by their appearance, humans tend to attribute those robots human-like abilities and, if their expectations fall short, the interaction may fail [14]. Humans strongly rely on implicit signals to cooperate; therefore, in this context, to obtain seamless human-robot collaboration, humanoid robots need to correctly interpret those implicit signals [3]. Furthermore, if we consider a scenario where the robot acts as helper or partner in an unconstrained environment, it acquires great importance to endow it with the ability of correctly estimating the characteristics of the handled objects; as a consequence, the robot can plan a safe and efficient motion action. In this study, we give particular attention to how a robot could assess an object features just by seeing it transported by a human partner. Inferring those properties from the human kinematics during the manipulation of the objects, rather than from their external appearance, grants the ability of generalizing over previously unseen items.

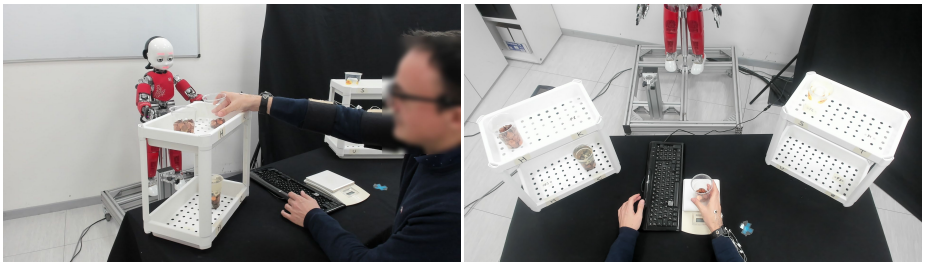
1.1 Rationale

Suppose to transport a glass full to the brim with water: the effort required to safely manage it without spilling a drop resembles the challenging scenario of porting an electronic device that could be damaged. If we want a robot to perform the same action, the first step would be to give the robot the capability of recognizing the intrinsic difficulty of the task; if we consider a hand-over task between a human and a robot, the latter should be aware that it is about to receive an object that requires a certain degree of carefulness in the handling. Moreover, an assessment of the weight of the object would allow an efficient lift. These features could be estimated from the human motion and ideally should be available before the end of the observed action, to trace the human abilities and to allow the robot to prepare for the possible handover. Differently from the weight, the concept of carefulness is not trivial. Previous studies have dealt with delicate objects, but focused more on robotic manipulation: the difficulty in the addressed tasks was given from the stiffness or the deformability of the item; tactile sensors where used for estimating the necessary force to apply a proper grasp [12, 16]. In our study we consider the carefulness necessary to move an item from a wider perspective. Indeed, not only the object stiffness but also its fragility, the content about to be spilled, or its sentimental value may lead a person to perform a particularly careful manipulation. In those real-life cases

we would like the robot to successfully estimate the carefulness required just by observing the human kinematics. As a proof of concept, we recorded some transportation movements involving glasses which differed for weight and carefulness levels and some kinematic features, derived from the human motion, were used to train classifier algorithms. These features were obtained for comparison both from a motion capture system and a robot camera. We hypothesize that, from the knowledge of kinematic features descriptive of the human movement: ***(H1)*** it is possible to infer if carefulness is required to manipulate an object and ***(H2)*** it is possible to discriminate a lighter object from a heavier one. To validate our hypothesis we have collected a dataset of human motions while performing a simple transporting task; then we have trained state-of-the-art classifiers to determine if it is possible to distinguish the carefulness associated with an object and its weight, exclusively observing the human motion.

2 Experimental setup

The experimental setup used to collect the data consisted of a table, a chair, two shelves (placed on different sides of the table) facing the volunteer, a scale, a keyboard with only one functioning key, and four plastic glasses (see Fig. 1).



(a) Lateral view (b) Top view

Fig. (1) Two views of the experimental setup with the volunteers in two grasp positions: on the shelf (1a) and on the scale (1b)

Table (1) Glasses features and abbreviations

Abbreviation	**Weight (gr)**	**Carefulness level**
W1C1	167	low (no water)
W2C1	667	low (no water)
W1C2	167	high (full of water)
W2C2	667	high (full of water)

The four glasses were characterized by two levels of associated carefulness and two weights, as shown in Table 1. The high level of carefulness was achieved filling the glass to the brim with water, while for the low level no water was placed in the glass. The different weights, instead, were obtained by inserting in the glasses a variable number of coins and screws; for the object with high level of carefulness the weight of the water was taken into account. Each object was weighted to guarantee a difference of 500 gr between light an heavy glasses. Glasses were identical in shape and appearance, and their transparency was chosen so that participants could clearly see the content of the glass and appropriately plan their movements . As displayed in Fig. 1, four positions were defined in each shelf, two on the top and two on the bottom level. These predefined positions were identified by a letter on a label.

Participants seated at the table and had to perform a structured series of reaching, lifting and transportation movements of the four glasses. The experiment started with all the four glasses on the shelves, the volunteer with their arms resting on the table and their right hand in the resting pose, marked with a blue cross (see Fig. 1b). During the experiment, the volunteers used their right hand to interact with the objects and their left to press the key of the keyboard. The experiment was structured as following:

- The volunteer pressed the key of the keyboard and a synthetic voice indicated the position on the shelf of the object to be transported. The position was referred to using the corresponding letter.
- The volunteer performed a reaching action toward the specified position and grasped the glass (see Fig. 1a).
- The volunteer performed a transportation action moving the glass from the shelf to the scale.
- The volunteer released the glass and returned to the resting pose.
- The volunteer pressed a second time the key and the synthetic voice indicated a position on the shelf where the glass should be transported. Of course, this time the selected position on the shelf was empty.
- The volunteer performed a reaching action towards the scale and grasped the glass (see Fig. 1b).
- The volunteer performed a transportation action moving the glass from the scale to the final position on the shelf.
- The volunteer released the glass and returned to the resting pose.

The participants repeated this sequence 8 times to familiarize with the task, while the main experiment consisted of 32 repetitions. A table containing the shelf initial and final poses for each repetition was previously designed to guarantee a good coverage of all the possible combinations of shelf positions and glasses. Each volunteer performed exactly the same experiment.
The experiment was conducted thanks to 15 healthy right-handed subjects that voluntarily agreed to participate into the data collection (7 females, age: 28.6 ± 3.9). All volunteers are members of our organization but none is directly involved in our research.

2.1 Sensors

The data used in this study was collected during the experiments previously described using a motion capture system from Optotrak, as ground truth, and one of the cameras of iCub. During the experiments other sensors have been used to collect data but their analysis is not in the scope of this paper. The humanoid robot iCub was placed opposite to the table and recorded the scene through its left camera, with a frame rate of 22 Hz and a resolution of the image of 340 x 240 pixels. The robot was just a passive observer and no interaction with the participants took place during the experiment. The Optotrak Certus®, NDI, motion capture (MoCap) system recorded the kinematic of the human motion through active infrared markers at a frequency of 100 Hz. The markers were placed on the right hand, wrist and arm. For the following analysis only a subset of the hand and wrist markers were considered (see Fig. 2). The data coming from the different sensors was synchronized through the middleware YARP [6] that gave to each sample a YARP timestamp. By pressing the key on the keyboard at the end of every trial the data coming from the MoCap were automatically segmented in different log files and the actual timestamp saved in a separate file. Successively the timestamps associated with the key pressures have been used to segment the data recorded by the robot camera.

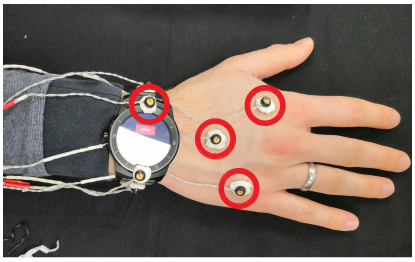


Fig. (2) Detail of the markers position on the right hand: those circled in red were interchangeably used to compute the features in each trial

Motion capture system data The data acquired by the motion capture system consisted in the tridimensional coordinates of each marker with respect to the reference coordinate frame of the Optotrak. Occlusions limited the MoCap visibility for specific part of the human movement. In our experiment the main source of occlusion was given by the presence of the shelves, in particular for the lower right positions. To partially overcome this problem, after a preliminary analysis, we chose to consider for each trial the most visible marker among a subset of four as representative of the movement. Indeed, during the transportation movements the hand could be assimilated to a rigid body. The four considered markers were placed respectively on the metacarpophalangeal joints of the index and of the little finger, on the diaphysis of the third metacarpal and

on the smartwatch in correspondence of the radial styloid (see the markers circled in red in Fig. 2 for reference). Two different interpolations, inpaintn [5] and interp1 of MATLAB ver. R2019b, have been used to reconstruct the data that are missing because of the occlusions. The data was filtered with a second order low pass Butterworth filter with a cutoff frequency of 10 Hz. Some trials have been excluded from the data set because of inconsistencies in the segmentation among the acquired sensors or because of errors of the subjects into pressing the key at the right moment, i.e. when their right hand was laying on the table in the resting position. Overall only 1.25% of the total acquired trials have been removed. Since our hypothesis is that it is possible to distinguish the features of the object that is being transported, it was necessary to isolate the transportation movement in every trial. To do so we took advantage of the experiment design. Indeed each trial presented three clearly identifiable phases: a reaching action, from the resting pose to the position occupied by the glass (either on the shelf or on the scale), a transportation movement and finally the departing (see Fig. 3). Our segmentation assumed that the start and end of the transportation phase is associated with a pick in the norm velocity of the hand. Therefore, the segmentation was performed by placing a threshold of 5% on the second peak of the norm of the velocity, after filtering it with a fourth order filter with a cutoff frequency of 5 Hz. The resulting data were then down-sampled to obtain the same frame rate as the camera of the robot.

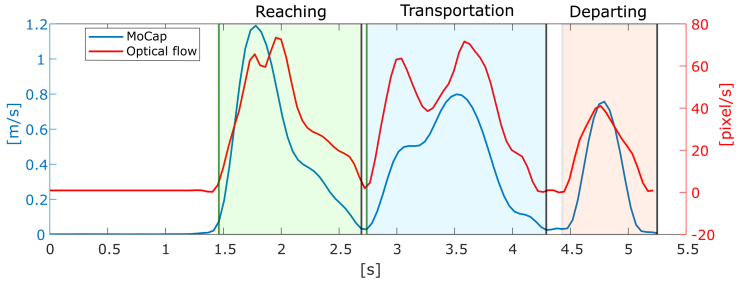


Fig. (3) Example of the velocity patterns from motion capture (in blue) and optical flow data (in red). The peaks characterizing the three phases of the trial (reaching, transportation and departing) are visible

Camera data and optical flow extraction As motion descriptor, from the saved raw images of the robot camera (see Fig. 4 for an example) we chose to compute the Optical Flow (OF), following an approach already tested [17, 18]. In this method, the optical flow is computed for every time instant using a dense approach [4], which estimates the apparent motion vector for each pixel of the image. The magnitude of the optical flow is thresholded to consider only those parts of the image where the change is significant. A temporal description of the motion happening in the derived region of interest is then computed averaging

the optical flow components. On the velocity extracted, a second order low-pass Butterworth filter with cutoff frequency of 4 Hz was applied to remove the noise (see Fig. 3).

(a) View from the iCub perspective (b) OF moving towards the right of the image (c) OF moving towards the left of the image

Fig. (4) Example of iCub view of the scene and the extracted OF. The colors codify for the direction of the movement: red is for motion towards the right part of the image (4b), blue for motion towards the left (4c)

3 Data pre-processing

The same set of motion representations was extracted during a pre-processing phase from both the motion capture data and the optical flow: the velocity $\mathbf{V}_i(t)$, the curvature $C_i(t)$, the radius of curvature $R_i(t)$ and the angular velocity $A_i(t)$ [9]. Their analytical expression is stated in Table 2. Such features can be computed for every time instant and by collecting them it is possible to progressively gather an increasing amount of information about the observed movement. This would then grant the robot the ability of discriminating online the characteristics of the object handled by the human partner. As shown in [17, 18], these data representations have been successfully used to discriminate online between biological and non-biological motion and to facilitate coordination in human-robot interaction [10]. In addition, kinematics properties, such as velocity, have been shown to be relevant in human perception of object weight [1]. Extracting those features during the pre-processing, instead of directly feeding the classification algorithms with raw data, allows to better compare the performance achieved with the two sources of data. Indeed, a precise control over the information used during the learning process is granted.

3.1 Dataset

As we have detailed before some sequences had to be removed for inconsistencies in the segmentation. This lead to a slightly unbalanced data set, containing more examples for specific classes. Indeed, class W1C1 had 235 sequences, class

Table (2) Motion features computed from motion capture and optical flow data

Motion feature	**Analytical expression**
Tangential velocity	$\mathbf{V}_i(t) = (u_i(t), v_i(t), \Delta_t)$
Tangential velocity magnitude	$V_i(t) = \sqrt{u_i(t)^2 + v_i(t)^2 + \Delta_t^2}$
Acceleration	$\mathbf{A}_i(t) = (u_i(t) - u_i(t-1), v_i(t) - v_i(t-1), 0)$
Curvature	$C_i(t) = \frac{\|\mathbf{V}_i(t) \times \mathbf{A}_i(t)\|}{\|\mathbf{V}_i(t))\|^3}$
Radius of curvature	$R_i(t) = \frac{1}{C_i(t)}$
Angular velocity	$A_i(t) = \frac{V_i(t)}{R_i(t)}$

W2C1 239, class W1C2 238 and class W2C2 had 236. Although cardinally the difference is minimum, to preserve the balance of the dataset we decided to fix the maximum number of sequences for each class to 235 and we have randomly selected the sequences for W2C1, W1C2 and W2C2. Notice that the four classes were characterized only by the weight and the carefulness level. Therefore other variables, such as the initial and final position of the glass and the direction of the movement, are not considered in the classification.

Due to the characteristics of the glasses, the duration of the transport movement varied consistently among the trials (i.e. the duration of the movement is consistently longer when the moved glass is full of water, belonging to the high carefulness class). To obtain sequences with the same number of samples for each trial, the segmented sequences were re-sampled, using the interp1 function of MATLAB. The number of samples was selected considering the class associated with the shorter duration of the transport phase, W1C1, and computing the median value among all its trials. The resulting value was 32. Therefore, our dataset was composed of two data structures: one derived from the MoCap data and the other one from the OF. Both structures had dimensions $940\,(trials) \times 32\,(frames) \times 4\,(features)$.

The re-sampling can be performed only knowing the start and end of the transportation phase. Since in an online scenario this information is not available, a further attempt was performed exploiting the ability of certain models to handle temporal sequences of different lengths. In this case, instead of re-sampling, a common zero-padding and masking technique were adopted. Therefore, the shorter temporal sequences were completed with zero values and those values were then ignored during the training, while the length of the longest transport movements was preserved. The shape of the data structures after the zero padding was: $940\,(trials) \times 132\,(frames) \times 4\,(features)$.

4 Classifiers

As introduced in Sect. 1.1, the goal of the classification is to discriminate between the two possible features of the transported glasses: ***(H1)*** the carefulness level associated with the object and ***(H2)*** the weight. Therefore, we decided to approach the problem using two binary classifiers, one for each feature, implemented in Python using Keras libraries [2]. As mentioned in Sect. 3.1 two models were tested: the first one relied on re-sampled features, while the second one used the original data with variable lengths.

4.1 Convolutional, Long-Short-Term-Memory and Deep Neural Network

Previous literature suggests that the combined use of Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN) is a good solution for classifying time dependent data, such as speech or motion kinematics [11, 7]. Therefore, our first model was inspired by [11] and consisted of two time distributed 1-D convolutional layers (that took as input 4 subsequences of 8 frames each), a max pooling and flatten layers, a 100 neurons LSTM, a 100 neurons Dense layer and a 2 neurons output layer with a sigmoidal activation function. A Leave-One-Out approach was adopted, to test the ability of the model to generalize over different participants. Thus, for each one of the 15 folds, the data from 14 subjects were used as training set and the data of the fifteenth participant as test set. The 20% of the data for each training set was kept for validation, and early stopping was implemented according to the validation loss function (with a patience parameter set to 5 epochs): this allowed to obtain good accuracy without incurring in overfitting. The batch size was fixed to 16. The model was fit with ADAM optimization algorithm and categorical cross-entropy as loss function. With respect to the model described in [11] some regularizers were added to avoid overfitting and make the network less sensitive to specific neurons weights. A L1-L2 kernel regularization was added to the two 1D convolutional layers ($l1 = 0.001, l2 = 0.002$) and a L2 kernel regularizer ($l2 = 0.001$) was added to the fully connected DNN layer; moreover, 0.5 dropouts were introduced.

4.2 Long-Short-Term-Memory and Deep Neural Network

The second model was implemented to test the possibility of generalizing over temporal sequences of variable length. To implement such an approach the data were padded with zeroes, as mentioned in the previous Section. Since the required masking layer was not supported by the Keras implementation of the CNN layer, we decided to opt for a simpler model: a 64 neurons LSTM, followed by a 32 neurons dense layer and a 2 neurons output layer with a sigmoidal activation function; also in this case L1-L2 regularization and 0.5% dropout were used to avoid overfitting. The optimization algorithm, the loss function and the validation approach with early stopping were the same as before. By using this

model the possibility of learning independently from the length of the temporal sequence was granted. This represents a further step towards the implementation of the same algorithm on the robot; indeed, no previous knowledge on the duration of the movement would be required to perform the classification, since the model is trained on variable temporal sequences.

5 Results

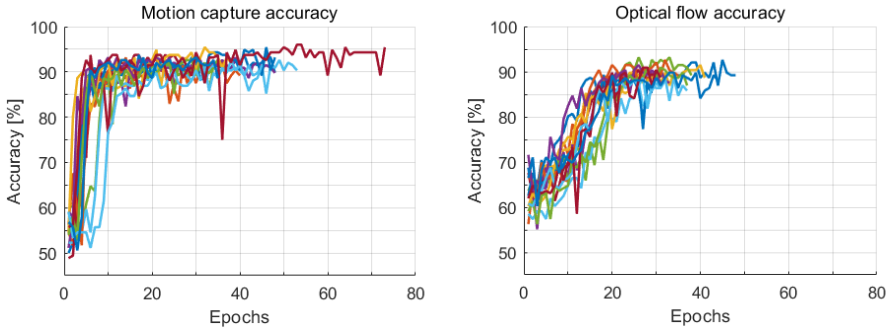
Results for the classifiers performance are presented for both the weight and the carefulness features and for both the considered source of data: the motion capture and the optical flow from the robot camera.

5.1 Carefulness level

The performances in the classification of the carefulness level with the model presented in Sect. 4.1 are reported in Table 3.

Table (3) Model accuracy (%, mean and standard deviation) on carefulness level classification with the CNN-LSTM-DNN model. In brackets are the results when volunteer 8 was included in the data set

	Motion capture	**Optical flow**
Training	92.15(92.00) ± 2.14(3.42)	94.03(92.18) ± 1.05(1.00)
Test	91.68(90.97) ± 5.00(11.12)	90.54(89.43) ± 6.56(7.59)



(a) MoCap accuracy (b) OF accuracy

Fig. (5) Accuracy in the carefulness classification with CNN-LSTM-DNN model for the validation set foe each fold. Accuracy from motion capture (5a) and from optical flow features (5b)

When performing the Leave-One-Out cross validation we noticed that the classification accuracy associated to volunteer 8 was significantly lower then the average (**MoCap test** *all:* 90.97 ± 11.12 *vol8:* 51.62; **OF test** *all:* 89.43 ± 7.59 *vol8:* 77.42). Examining the experiment videos we noticed that the volunteer 8 was particularly careful even when handling the glasses not containing water. Our impression was confirmed after computing the median duration of the not careful movements among the subjects. The duration for volunteer 8 (2.04±0.18 seconds, median and median absolute deviation) differed significantly from the ones of the other participants, as for the rest of the group the median duration resulted 1.47 ± 0.15 seconds (Kruskal-Wallis test: $\chi^2(14, N = 480) = 136.8, p < .01$). In Table 3 we have reported in brackets the results when including this subject in the dataset. As can be observed, when the participant was included the accuracy variance on the test increased significantly for each the sensing modalities.
Figure 5 shows the trend of the accuracy over the epochs for the validation set of each one of the folds. Comparing the graphs for the two sources of data ((a) motion capture, (b) optical flow) it can be noticed how the first one reaches an accuracy above the 80% in less than 10 epochs, while, using the features from the optical flow, more training is necessary to reach the same level of accuracy (over 20 epochs). Furthermore, the accuracy trend of the motion capture features is more stable.

Similarly, the carefulness classification performance with the model presented in Sect. 4.2, fed with the original temporal sequences of variable lengths, is shown in Table 4. As before, the variability in the test accuracy reduced when volunteer 8 is excluded from the dataset, and the overall accuracy improved for both the sensing modalities. With this model, compared to the values in Table 3, the accuracy achieved with the MoCap data is higher, while the one of the OF slightly reduced.

Table (4) Model accuracy (%, mean and standard deviation) on carefulness level classification for simpler LSTM-DNN model. In brackets the results considering volunteer 8

	Motion capture	**Optical flow**
Training	96.57(94.32) ± 1.19(1.77)	92.10(90.39) ± 4.58(2.56)
Test	95.17(92.66) ± 5.56(8.49)	88.38(86.50) ± 8.68(10.75)

5.2 Weight

In Table 5 are shown the results for the classification of the weight achieved with re-sampled data on the first implemented model. In this case, volunteer 8 did not present any peculiarity and therefore it was included in the dataset. As we can observe in Table 5, the accuracy with the motion capture data is above 60% and is higher than the one obtained from the optical flow.

Table (5) Model accuracy (%, mean and standard deviation) on weight classification with the CNN-LSTM-DNN model, fed with re-sampled data

	Motion capture	**Optical flow**
Training	64.10 ± 2.34	55.24 ± 2.37
Test	61.83 ± 7.16	54.47 ± 4.29

Finally, Table 6 reports the accuracy for the weight classification with the LSTM-DNN model, fed with the original temporal sequences of different lengths. In this case the performance was comparable between the data from the two sensing modalities.

Table (6) Model accuracy (%, mean and standard deviation) on weight level classification for the second model, LSTM-DNN

	Motion capture	**Optical flow**
Training	54.95 ± 2.66	55.30 ± 1.95
Test	54.75 ± 5.27	53.29 ± 3.59

We have noticed that, despite adopting the same approach, the accuracy on the weight classification is not as satisfying as the one achieved for the carefulness. A possible explanation of these results could be related to the different effect that weight may have on different transport movements. Possibly the weight influence varies if the transportation is from top to bottom or vice-versa. Furthermore, the presence of water in some of the glasses may have led the subjects to focus mainly on the carefulness feature, unconsciously overlooking the weight difference. Therefore, we add two specifications of the second hypothesis: ***(H2.1)*** the influence of the weight during transportation is dependent on the trajectory of the motion; ***(H2.2)*** when an object is associated with an high level of carefulness, the weight has a limited influence on the transportation movement. Both hypotheses were tested with the first model, which gave better results for the weight classification. Concerning the first hypothesis, we reduced the variability in the movements and tried to discriminate the weight in the subset of transport movements from the scale towards the shelves (**MoCap:** *Tr*: 68.90 ± 2.68 *Test*: 63.42 ± 8.96; **OF:** *Tr*: 59.10 ± 4.27 *Test*: 55.17 ± 6.24); there is a slight improvement for both the data sources compared to the values in Table 5. Notice that the trajectories still have a discrete amount of variability since the position to reach on the shelf could be left or right, high or low. The second hypothesis was investigated by testing the weight discrimination within the subset of objects which required the same carefulness level: low (**MoCap:** *Tr*: 64.49 ± 5.24 *Test*: 61.93 ± 6.86; **OF:** *Tr*: 62.52 ± 3.53 *Test*: 56.84 ± 6.77) or high (**MoCap:** *Tr*: 62.72 ± 3.65 *Test*: 59.03 ± 8.73; **OF:** *Tr*: 57.92 ± 1.31 *Test*: 53.48 ± 7.63).

For both the tests the results are inconclusive, since the classification accuracies have not changed much respect to the ones reported in Table 5. It should be noted though that the dimension of the dataset used to validate hypotheses ***(H2.1)*** and ***(H2.2)*** halved, which has an impact on the statistical relevance of the results.

6 Discussion

Regarding the carefulness feature, as reported in Table 3 the first classifier is able to correctly discriminate if the transportation of the object requires carefulness or not, independently from the sensing modality used. Considering the performance on the data coming from the two sources, no significant difference is detected between them. Therefore, not only using an accurate system such as the motion capture, that integrates sensory inputs from different locations to estimate the position in space of the target, but also using the camera of the robot (single point of view), it is possible to extract features to discriminate between careful and not careful motions. Figure 5 shows an insight on how the learning process advanced for the two data sources. Even though the final performances are comparable, it can be appreciated how the model trained with the features from the motion capture converges quicker to an accuracy value above the 80%.
The approach adopted with the second classifier is more general, in the sense that data are not re-sampled to have the same dimension but the variability in their duration is taken into account. Even though this model is simpler, with just one LSTM and one dense layer, the performance on the carefulness classification considering the MoCap data increased (see Table 4 for reference). Although the accuracy using the optical flow is slightly lower, we consider this as a promising step towards the implementation of the same algorithm on the robot.

Concerning the weight, the accuracy achieved for both the sensing modalities and for both the models is lower than the one obtained for the carefulness (see Tables 5 and 6 for reference). To explain this outcome in Sect. 5.2 we have formalized two additional hypotheses. ***(H2.1)*** was inspired by [8], where it has been proposed that the vertical component of the velocity during the manipulation of an object is perceived by humans as informative about its weight. Since the trials in our dataset explored a variety of directions and elevations, this introduced a great variability in the vertical component of the velocity. Instead, concerning ***(H2.2)***, we have supposed that the greatest challenge for the volunteers during the experiment is to safely handle the glasses full of water; the difference in weight between the objects was not remarkable in comparison with the stark contrast given by the presence (or absence) of water. As mentioned in Sect. 5.2 the first classifier was tested against these hypotheses, but no significant improvements in the accuracy have been achieved. Given the results of our experiment we can not validate hypothesis ***(H2)***. However, since we have explored only a subset of the possible kinematic features we can not argue against this hypothesis either. A possibility for future works is to focus on the

vertical component of the velocity. Furthermore, ***(H2.1)*** and ***(H2.2)*** should be explored on reasonably extended datasets to obtain more reliable results.

7 Conclusions

As human-robot interactions are becoming increasingly frequent, it is crucial that robots gain certain abilities, such as the correct interpretation of implicit signals associated with the human movement. In this study we focused on two fundamental implicit signals commonly communicated in human movements: the impact of the weight and the carefulness required in the object manipulation (e.g. transport of fragile, heavy and unstable objects). Our hypotheses aimed to demonstrate that it is possible to discriminate between lighter and heavier items ***(H2)*** and to infer the carefulness required by human operator in manipulating objects ***(H1)***. We proved that it is feasible to reliably discriminate when the human operator recruits motor commands of careful manipulation during the transportation of an object. Indeed, it is reliable to estimate extreme carefulness from two different typologies of sensory acquisition: from motion tracking system and from the single view point of the robot‘s camera observing the movement. On the other hand, the proposed algorithms show lower accuracy when applied to weight classification, and these results does not allow us to validate our second hypothesis. The estimation of the weight from human motion should be subject of further studies, exploring other classification strategies or kinematic features subset (e.g. extraction of the vertical components of the velocity during manipulation). This study firmly supports the research in human-robot interaction, especially in the direction of addressing complex situations in realistic settings (e.g.: industrial environment, construction site, home care assistance, etc.). In these specific scenarios the robot can autonomously leverage on insights inferred from implicit signals, such as the carefulness required to move a object, in order to facilitate the cooperation with the human partner.

Acknowledgement

This paper is supported by European Union's Horizon 2020 research and innovation program under grant agreement No 870142, project APRIL (multi-purpose robotics for mAniPulation of defoRmable materIaLs in manufacturing processes).

References

1. Bingham, G.: Kinematic form and scaling: further investigations on the visual perception of lifted weight. Journal of experimental psychology. Human perception and performance **13**(2), 155—177 (1987)
2. Chollet, F., et al.: Keras. https://keras.io (2015)

3. Dragan, A.D., Lee, K.C.T., Srinivasa, S.S.: Legibility and predictability of robot motion. In: Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction, pp. 301–308. Tokyo, Japan (2013)
4. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: In: Proceedings of the 13th Scandinavian Conference on Image Analysis, LNCS 2749, pp. 363–370. Gothenburg, Sweden (2003)
5. Garcia, D.: Robust smoothing of gridded data in one and higher dimensions with missing values. Computational Statistics and Data Analysis **54**, 1167–1178 (2010)
6. Metta, G., Fitzpatrick, P., Natale, L.: Yarp: Yet another robot platform. International Journal of Advanced Robotic Systems **3** (2006)
7. Neverova, N., Wolf, C., Lacey, G., Fridman, L., Chandra, D., Barbello, B., Taylor, G.: Learning human identity from motion patterns. IEEE Access **4**, 1810–1820 (2016)
8. Palinko, O., Sciutti, A., Patane, L., Rea, F., Nori, F., Sandini, G.: Communicative lifting actions in human-humanoid interaction. In: Proceedings of the 14th IEEE-RAS International Conference on Humanoid Robots, pp. 1116–1121. Madrid, Spain (2015)
9. Press, C.: Action observation and robotic agents: Learning and anthropomorphism. Neuroscience Biobehavioral Reviews **35**(6), 1410 – 1418 (2011)
10. Rea, F., Vignolo, A., Sciutti, A., Noceti, N.: Human motion understanding for selecting action timing in collaborative human-robot interaction. Frontiers in Robotics and AI **6**, 58 (2019)
11. Sainath, T.N., Vinyals, O., Senior, A., Sak, H.: Convolutional, long short-term memory, fully connected deep neural networks. In: Proceedings of the 2015 IEEE ICASSP, pp. 4580–4584. Brisbane, Australia (2015)
12. Sanchez, J., Corrales, J.A., Bouzgarrou, B.C., Mezouar, Y.: Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey. The International Journal of Robotics Research **37**(7), 688–716 (2018)
13. Sanchez-Matilla, R., Chatzilygeroudis, K., Modas, A., Duarte, N.F., Xompero, A., Frossard, P., Billard, A., Cavallaro, A.: Benchmark for human-to-robot handovers of unseen containers with unknown filling. IEEE Robotics and Automation Letters **5**(2), 1642–1649 (2020)
14. Sandini, G., Sciutti, A.: Humane robots—from robots with a humanoid body to robots with an anthropomorphic mind. ACM Transactions on Human-Robot Interaction **7**, 1–4 (2018)
15. Sciutti, A., Patane, L., Nori, F., Sandini, G.: Understanding object weight from human and humanoid lifting actions. Autonomous Mental Development, IEEE Transactions **6**, 80–92 (2014)
16. Su, Z., Hausman, K., Chebotar, Y., Molchanov, A., Loeb, G.E., Sukhatme, G.S., Schaal, S.: Force estimation and slip detection/classification for grip control using a biomimetic tactile sensor. In: Proceedings of 15th IEEE-RAS International Conference on Humanoid Robots, pp. 297–303. Seoul, Korea (2015)
17. Vignolo, A., Noceti, N., Rea, F., Sciutti, A., Odone, F., Sandini, G.: Detecting biological motion for human–robot interaction: A link between perception and action. Frontiers in Robotics and AI **4**, 14 (2017)
18. Vignolo, A., Rea, F., Noceti, N., Sciutti, A., Odone, F., Sandini, G.: Biological movement detector enhances the attentive skills of humanoid robot icub. In: Proceedings of the 16th IEEE-RAS International Conference on Humanoid Robots. Cancun, Mexico (2016)