Skeleton-based Action and Gesture Recognition for Human-Robot Collaboration

Matteo Terreran, Margherita Lazzaretto, and Stefano Ghidoni

Dept. of Information Engineering, University of Padova, Padova, Italy
[matteo.terreran, stefano.ghidoni]@unipd.it
margherita.lazzaretto@studenti.unipd.it

Abstract. Human action recognition plays a major role in enabling an effective and safe collaboration between humans and robots. Considering for example a collaborative assembly task, the human worker can use gestures to communicate with the robot while the robot can exploit the recognized actions to anticipate the next steps in the assembly process, improving safety and the overall productivity. In this work, we propose a novel framework for human action recognition based on 3D pose estimation and ensemble techniques. In such framework, we first estimate the 3D coordinates of the human hands and body joints by means of OpenPose and RGB-D data. The estimated joints are then fed to a set of graph convolutional networks derived from Shift-GCN, one network for each set of joints (i.e., body, left hand and right hand). Finally, using an ensemble approach we average the output scores of all the networks to predict the final human action. The proposed framework was evaluated on a dedicated dataset, named IAS-Lab Collaborative HAR dataset, which includes both actions and gestures commonly used in human-robot collaboration tasks. The experimental results demonstrated how the ensemble of the different action recognition models helps improving the accuracy and the robustness of the overall system.

Keywords: human action recognition, gesture recognition, 3D pose estimation, ensemble learning, human-robot collaboration

1 Introduction

Human-robot collaboration (HRC) aims to a close and direct collaboration between robots and humans to reach higher productivity and ergonomics thanks to the synergy between human intelligence and robot mechanical power [1–3]. In such scenario, the robot should always be aware of the position and the intentions of the human worker, in order to foresee possible dangerous situations and guarantee the safety of the human partner. Moreover, by knowing which step of the assembly the person is working on, the robot can plan its actions accordingly, such as working in a different part of the workspace or preparing parts and tools for the next step in the assembly process.

Human action recognition has been widely investigated in the literature to provide such awareness to the robot [4–6]. Generally, the actions to be recognised

are steps in an assembly sequence (e.g. pick a part, place a part, screw) or general actions such as walking, interacting or standing still; thus all actions that involve various parts of the body and can be easily distinguished. Some works focus instead on gesture recognition, where the actions to be classified are usually very small movements of few parts of the body, such as hands, used to communicate information to the robot (e.g. go left, go right, stop). Given this difference, the problems of action and gesture recognition are generally tackled separately with dedicated approaches and setups, for example a camera framing only the human hands in case of gesture recognition.

In this work, we address the problem of human action recognition in collaborative scenarios by proposing a general framework capable of recognizing both hands gestures and full-body gestures (i.e., general actions involving the whole body). Our framework relies on an ensemble of different skeleton-based classifiers, each one trained to recognize actions on a subset of skeleton joints (e.g., body joints, hands joints); the use of dedicated classifiers allows to be robust to partial skeleton inputs with missing joints, and to handle together body actions and hand gestures. Moreover, using skeletons as an intermediate representation between the classifiers and the raw RGB-D data, our system is able to generalize to different scenarios and collaborative tasks. Indeed, in such a manner, the action classifiers can leverage on mid-level features describing only people movements: picking up a piece to be assembled or grabbing up a hammer from a toolbox can both be considered a "pick" action despite the particular object being picked up.

The proposed system has been evaluated on a dataset acquired on purpose in our laboratory, namely the IAS-LAB Collaborative HAR dataset[1], which includes RGB-D videos of several subjects executing typical actions which occur in a human-robot collaboration task. Unlike other works in the literature, in our dataset we did not restrict the action set to a specific application (e.g., take part A, place part B, take hammer from the toolbox), but we tried to generalize the most recurrent actions and gestures in the literature in order to create a common basis for several collaborative tasks.

Summarizing, the work presents 3 main contributions: (i) a unified framework for human action and gesture recognition in a human-robot collaboration scenario; (ii) an experimental comparison of different ensembling techniques to improve the overall accuracy and robustness of the system; (iii) a novel RGB-D dateset for action recognition in a human-robot collaboration scenario, including both general actions and hand gestures, to further drive research in this field. The remainder of the paper is organized as follows. Section 2 reviews the works related to action recognition, with a focus on human-robot collaboration scenarios. In Section 3 the main elements of our system are described in details. In Section 4 we present the action recognition dataset acquired in our laboratory, used to thoroughly evaluate the system proposed in Section 5. Finally, in Section 6, conclusions are drawn and future directions of research identified.

[1] Available at http://robotics.dei.unipd.it/

2 Related Works

Human action recognition (HAR) is generally defined as the process of identifying and analyzing the movements of one or several parts of the human body, applied in a large set of scenarios such as public events [7], video surveillance and home monitoring [8, 9], human-robot interaction [5, 10] and safety monitoring in industry [11]. According to the sensory modalities and sources adopted, HAR systems can be classified into two main categories: contact-based and vision-based systems. Contact-based approaches require a physical interaction of the user with sensors and acquisition devices such as accelerometers, multi-touch screens, body-mounted sensors or wearable sensors like sensorized gloves [9, 10, 12]. Vision-based approaches rely instead on images or video sequences to recognize activities, using either a single RGB-D camera or a network of cameras to be robust to occlusions. Unlike contact-based activity recognition systems, vision-based systems are considered "non-intrusive" since they do not require ordinary users to wear several and uncomfortable devices on different parts of their body, making them an easier solution to be used in real scenarios.

A main challenge in vision-based action recognition tasks consists in handling both a spatial and a temporal dimension, since an action is generally seen as a sequence of consecutive movements in time. Many approaches have been proposed in the literature to deal with it, such as LSTMs [13, 14], 3D-CNNs [15, 16] and multi-stream 2D-CNNs [17, 18]. In 3D-CNNs the temporal information is exploited by considering an input sequence of RGB frames, then elaborated using 3D convolution kernels. In multi-stream CNNs instead, spatial and temporal information are analyzed by two different branches in the network, taking as input RGB frames and optical flow information respectively. Recently, many pose estimation models such as OpenPose [19] achieved very high performance and many researchers started addressing action classification using the human body pose as input of graph convolutional networks (GCNs) [20–22]. Body pose allows to encode spatial and temporal information in a more compact way than images, leading GCN models such as the Shift-GCN architecture [20] to outperform other approaches on many popular action recognition datasets [23, 24].

Human action recognition is largely applied in many human-robot interaction scenarios, such as social robotics and manufacturing industry. Usually, the set of actions to be recognized includes hand gestures and face expressions, in order to provide a fast and easy interaction with the robots [4, 25]. For example, in [4] authors propose six gestures to allow the human to communicate with a collaborative robot; the gestures (i.e., *on*, *off*, *right*, *left*, *up*, *down*) are recognised by fusing together three different modalities, namely speech command (with a CNN), hand motion (with a LSTM), and body motion.

Considering an industrial context, the set of actions of interest can also include either general actions (e.g., walking, standing) or very specific actions and gestures, depending on the main purpose of the action recognition. For example, action recognition could be used to guarantee human safety in human-robot collaboration by monitoring what people do while inside the robot workspace. In [5], authors monitor the people moving close to a robotic arm and recognize

actions such as *passing*, *observing*, *dangerously observing*, *interacting*. The recognition is done with a multi-modal approach: a 3D-CNN is used to learn features from sequences of RGB frames, while signals collected with a haptic sensors are employed to detect collisions between human and robot by means of a 1D-CNN.

When a task involves close collaboration between humans and robots, human action recognition could be used to improve productivity, by monitoring the different stages of the collaboration. In such case the set of actions to be recognized generally includes the different operations that the human needs to fulfill the overall assembly task. For example, in [10] the set of actions includes *grab a tool* from a toolbox, *insert a screw*, *tight the screw*, *put back the tool* in the toolbox. In such work, authors proposed an action recognition classifier based on a CNN trained on a combination of skeleton features and signals from EMG and IMU sensors. In [6] similar activities are considered, such as *taking a product or a component*, *move a product*, *grab a tool*, *put on screws*, *hold a product*, *tighten the screws*, *check product* and *place product*. To classify such actions the authors focus on hands information only, proposing a system that combines hands' pose and images cropped around the hands. In [17] the assembly actions include instead *cleaning*, *hammering*, *polishing*, *smearing*, *installing*, *screwing* and *marking*, all recognised using a two-stream CNN.

Despite sharing some common high-level actions and gestures, many of the human-robot collaboration systems presented in this section consider a very specific set of actions related to a particular task; in many cases authors acquire also an ad-hoc dataset for their application which, however, is either not published or proves to be too specific to be used by other researchers. Moreover, action recognition is usually addressed focusing either on body information or hands information. To the best of our knowledge no work has attempted to recognise actions using body and hands information together.

In this work we address both such problems, namely the lack of a general dataset and the recognition of actions involving either the body or the hands. On one hand, we propose a general framework to recognize both body actions and hands gestures in a collaborative scenario; on the other hand the system is evaluated on a novel action recognition dataset acquired on purpose, including common actions of a collaborative task so as to be generalisable to various applications.

3 Methods

In this section, we provide a detailed description of the main parts of our system. A schematic representation of the proposed pipeline is shown in Figure 1, highlighting the main steps involved. Our system takes as input a sequence of RGB-D frames and predicts the corresponding action performed according to a given set of actions of interest. In the first stage, we estimate the pose of the person in a sequence of RGB frames by means of OpenPose [19]; in particular, we focus on estimating the pose information of both the body and the hands, since many collaborative gestures could be executed by using the hands only.

OpenPose predicts a set of 2D joints describing the pose of the person in the image, which is then projected in 3D space thanks to the information provided by the sequence of depth images. The estimated 3D joints are then fed to a set of graph convolutional networks derived from Shift-GCN [20], a state-of-the-art action recognition architecture. Indeed, for the final action recognition stage, we rely on an ensemble of classifiers, each one trained to recognize actions from a different set of joints (e.g., body, left hand and right hand). The final output of the system is a weighted average of the output scores of all the classifiers, which according to our experiments in Section 5 proves to be more accurate and robust with respect to a single model trained on body and hands poses together.

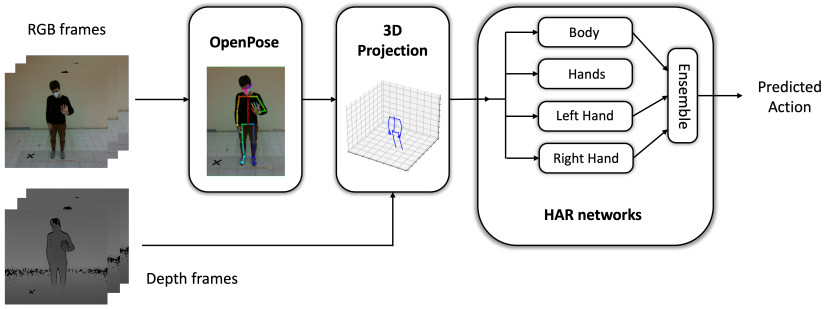

Fig. 1: Overview of the proposed action recognition system. In the first stages, 3D skeletons including body and hand joints are estimated from sequences of RGB-D frames by means of OpenPose architecture. In the last stage, the classification step, actions are predicted using an ensemble of skeleton-based action recognition models, each one trained to recognize actions from a particular set of 3D joints.

3.1 Pose estimation

To develop our general framework we focused on skeleton-based action recognition models, which are in general more accurate and robust than other approaches. Indeed, skeletons present many advantages with respect to raw RGB-D data, since they represent a lightweight information describing both spatial and temporal information in a compact form (i.e., joints' coordinates). Moreover, skeletons provide a robust representation of the human movements free of any disturbances such as external objects, lighting and aesthetic differences of people such as clothes or skin colour; this is important especially for collaborative scenarios, where both the human and the robot are moving and the human worker should interact with many objects and tools.

For pose estimation in our framework we rely on the OpenPose architecture [19], which provides different pretrained models for multi-person pose estimation, allowing to estimate in real-time either 15, 18 or 25 body keypoints, 42 hand keypoints and 70 face keypoints. Although very accurate in general,

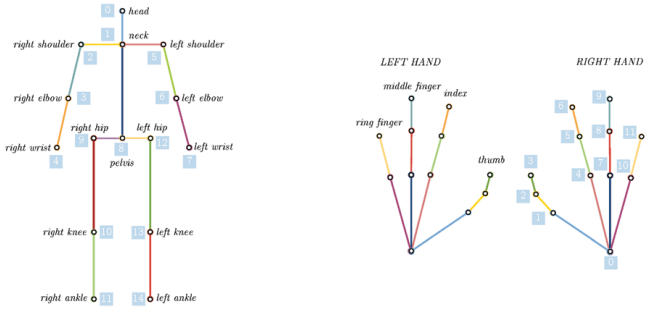

Fig. 2: Body and hand joints considered in the proposed framework, together with their associated ID. We consider a total of 15 joints describing the body, and 12 joints to describe each one of the hands.

OpenPose predictions could be missing of some keypoints when the subject is far away from the camera or partially occluded; this is especially true for the hands' keypoints, which can be easily occluded during interactions with objects.

Considering such limitations, in our framework we consider as input of the HAR models only a subset of the most significant joints for the body and hands skeletons provided by OpenPose. In particular, as shown in Figure 2, our choice for the hand joints is limited to: the wrist, three joints each for the thumb, index and middle finger, and two joints for the ring finger; for a total of 12 joints for each hand. The rest of the joints are omitted, being more difficult to be estimated and less useful for the action recognition process. For what concerns the body, some keypoints such as the ones corresponding to eyes, ears and feet are not considered either, leading to a 15 joints body model comprising head, neck, shoulders, elbows, wrists, pelvis, hips, knees and ankles.

3.2 3D Pose estimation

The output of OpenPose is a 2D skeleton describing the pose of each person in the input image. From this information we compute a 3D skeleton by means of re-projection, in order to have a more standard and general representation to be used as input for the action classifiers. Given a calibrated RGB-D camera and its extrinsic parameters, it is possible to transform the depth image by aligning it to the RGB image, obtaining a direct mapping for each pixel; for each 2D keypoint (x_P, y_P) estimated by OpenPose in the input image, we use the information in the aligned depth image to compute the keypoints coordinates in the 3D space using re-projection and the intrinsics parameters of the camera.

The depth information acquired with a RGB-D sensors is usually not accurate around borders and for small objects, like the fingers of the hand in our case. Therefore, to improve accuracy and robustness of the 3D keypoints, we do not consider the raw depth value, but we take instead the median value in a 5×5 window centered at coordinates (x_P, y_P) in the depth image.

3.3 Action and gesture recognition

The action recognition module in our framework is based on the Shift-GCN architecture [20], a graph convolutional network which achieved state-of-the-art performance on the NTU RGB+D dataset [23]. Compared to other architectures, Shift-GCN has the advantage of being lightweight and requiring a lower computational cost, thanks to the introduction of the graph shift convolution operations. This was one of the main aspects that guided our choice, since in human-robot collaboration scenarios we would like to have a fast recognition of the human actions to obtain a smooth and responsive collaboration.

The Shift-GCN architecture takes as input a sequence of skeletons, (i.e., a sequence of 3D joints coordinates), and is composed of 10 blocks, each one including a spatial graph shift convolution operation and an adaptive temporal shift operation. The final output layer contains 60 nodes, since it has been originally proposed to classify the actions in the NTU RGB+D dataset.

Unlike the NTU dataset, where actions were classified based only on the body pose information, in this work we aim to develop a general framework capable of recognizing both actions and gestures which can occur during the collaboration between a human worker and a robot; in such scenario, many possible actions and gestures involve only the use of hands while the body of the worker stands still (e.g., "stop" and "confirm" signals). This leads us to propose different action recognition networks based on the Shift-GCN architecture, each one designed to recognize collaborative actions from a particular set of joints. In particular, we consider the following set of joints:

- **wholebody**, which includes the 39 joints shown in Figure 2 describing the pose of both the body and the hands;
- **body**, including the 15 joints which describe the pose of the body;
- **hands**, including the 24 joints which describe both hands together;
- **single hand**, including the 12 joints of a hand (i.e., wrist and fingers' joints).

As in the original Shift-GCN architecture, the input sequences of skeleton joints undergo some pre-processing operations before being fed to the network: a traslation of the joint coordinate reference frame to a central joint of the skeleton, and a joints coordinate normalization. These operations allow to consider all the body movements with respect to the body itself, which makes them a better input for the network and easier to generalize to different scenarios. For our networks, we choose as origin of the new reference frame the joint corresponding to the neck (i.e., Joint 1 in Figure 2); the $z-$axis of the new reference frame is taken parallel to the segment connecting the pelvis (i.e., Joint 8) and the neck joints, while the $x-$axis is considered parallel to segment connecting the shoulder joints (i.e., Joints 2 and 5). For the networks considering only the hands joints we use the same convention for the reference frame's axes, but placing its origin on the wrist joint (i.e., Joint 0) of each hand (or the right hand wrist when both hands are considered. In such a manner, also the hands movements are expressed with respect to a local reference frame maintaining their relative orientation with respect to the rest of the body.

3.4 Ensemble averaging of the classifiers predictions

As the last step of our proposed framework, all the action recognition models' outputs are combined together by means of ensemble techniques to compute the final prediction. Ensemble is a machine learning technique that combines several base models in order to produce one optimal predictive model, improving the overall accuracy and robustness.

We propose two main approaches to combine the information from the action recognition models: an ensemble of the *body* and *hands* models, and an ensemble of the *body* model with both single hand models (i.e., *left_hand* and *right_hand* models). In both approaches, the information from the models is combined at the score-level, namely the output of the *softmax* activation function in the last layer of the networks. Considering for example the first approach (i.e., *body* + *hand* models), we compute the final score as a weighted sum of the score of each model. The predicted action l_{pred} is then obtained by taking the argmax of the final score,

$$l_{pred} = arg\,max\left(\alpha_b\,\mathbf{o}_b + \alpha_h\,\mathbf{o}_h\right), \quad with\ \alpha_b + \alpha_h = 1 \tag{1}$$

where N is the number of actions of interest, $\mathbf{o}_b \in [0,1]^N$ is the output score of the *body* network, $\mathbf{o}_h \in [0,1]^N$ is the output score of the *hands* network and α_o, α_h are the corresponding weights.

In the second approach (i.e., *body* + *left_hand* + *right_hand* models) we take into account also the fact that some actions or gestures can be made using only one hand (e.g., confirm, left, right, stop) while the other one remains still or even not visible. The final score is still computed as a weighed sum of the models' score, but considering a weight $\alpha_{ih} = 0$ if the hand $i, i \in \{left, right\}$ is the only one not visible or in a rest position. We assume in this case that the hand whose action is labelled as "rest" falls in a group of actions in which only one of the two hands is actively moving, while the other one stands still; the actual action is therefore related to the moving hand and the other one is irrelevant in terms of action recognition.

4 IAS-Lab Collaborative HAR Dataset

As highlighted in Section 2, there exist different actions that are common to many human-robot applications. In order to obtain a general framework suitable for different collaborative settings, we trained our models on a set of actions and gestures that generalize those commonly found in the literature. For example, in many works we have actions such as *grab a tool*, *pick a piece* or *pick the hammer* which can be all summarized as a *pick* action if we only focus on the human movements; the specific object to be picked can be then detected from a dedicated object detector. The actions that we selected are reported in Table 1, categorized in four main groups. The first group refers to general movements that a person can do in the robot workspace, where *WALK* includes all the movements of a worker moving around the workspace and *REST* indicates that the human operator is not working. The second group includes the most common

Table 1: Actions and gestures in the IAS-Lab Collabative HAR dataset

Group	Actions
Spatial movements	WALK, REST
Assembly actions	PICK, PLACE, SCREW, INSERT/JOIN, HAMMER
Collaborative gestures	HAND TO, REQUIRE
Communication gestures	STOP, OK/CONFIRM, UP, DOWN, FORWARD, BACKWARD, LEFT, RIGHT, POINT

assembly actions performed by the human worker, while the third group includes all the signals that the human can do to request or pass objects to the robot during the collaboration. Finally, the fourth group includes all the gestures used to communicate an instruction to the robot, including directions of movements and signals of confirmation or halting.

The set of actions to be recognised was chosen in order to include movements that are as generic as possible, and that can in general be distinguished without pairing them to objects or tools and without setting a world coordinates system to locate the human operator in the workspace. It can be noticed that many actions require an active use of hands and that the hands movements are potentially discriminative in the recognition of the performed action, especially for what concerns gestures and assembly actions.

Once defined the set of actions in Table 1, we collected a novel dataset of people performing such actions in our laboratory. We asked to 6 different subjects to perform 5 times each the 18 actions selected, for a total of $18 \times 6 \times 5 = 540$ samples. Each sample is a sequence of RGB-D frames of about 5 seconds, acquired with a Intel Realsense L515 camera[2]. The camera was placed at a distance of about 2.5 meters from the subjects, in order to frame the whole subject body during all the actions. Some samples from the acquired dataset are shown in Figure 3. During the dataset collection, subjects were only told the name of the action to be executed, without receiving any further instruction on how to perform it. This allowed to increase the variability of the dataset, since many subjects performed the same actions in different manners as shown in Figure 4. Indeed, we aim to recognise actions and gestures executed as naturally as possible, without creating a rigid dictionary of movements and having to give many specific instructions to users. This can help achieving an easier, more immediate and natural communication between people and robots.

[2] https://www.intelrealsense.com/lidar-camera-l515/

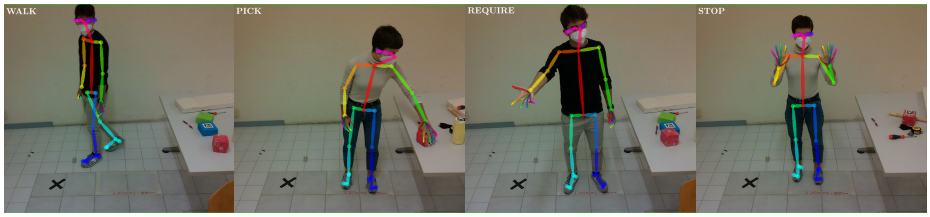

Fig. 3: Some samples from the IAS-Lab Collaborative HAR dataset, together with the skeletons outputs estimated by means of OpenPose.

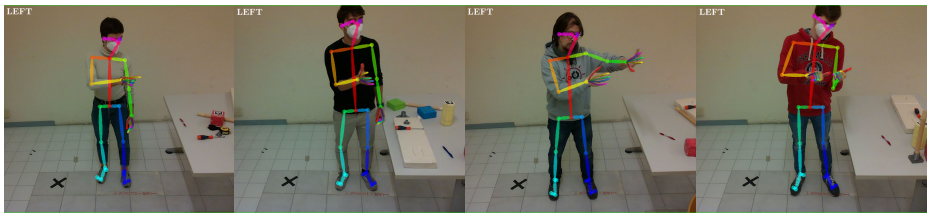

Fig. 4: Examples of variability in the IAS-Lab Collaborative HAR dataset. Each subject performed the actions requested in a different manner.

5 Experimental Results

The proposed framework has been evaluated using the IAS-Lab Collaborative HAR dataset described in Section 4. Such dataset is very small with respect to the dataset commonly used to train deep learning models, and hence not suitable to train from scratch the action recognition networks proposed in our framework. Therefore, the action recognition models have all been trained on a large collection of data such as the NTU RGB+D dataset and then fine-tuned on the IAS-Lab Collaborative dataset. This allows to exploit the larger dataset to learn several features encoding the human movements, which are then specialized for the human-robot collaboration scenario. The models described in the following have all been implemented starting from the official implementation of the Shift-GCN architecture[3]. In all the experiments the models have been trained using the hyper-parameters suggested by the Shift-GCN authors, using a NVIDIA® Titan RTX 2080 GPU.

5.1 Pre-training on the NTU RGB+D dataset

The NTU RGB+D dataset [23] is a large dataset containing RGB-D frames and pose annotations of several subjects performing 60 actions, acquired with a multi-view setup composed of 3 cameras. Pose annotations are provided as skeleton models composed of 25 joints describing only the body pose. However,

[3] https://github.com/kchengiva/Shift-GCN

they do not include hands joints and the body joints slightly differ from the ones estimated by the OpenPose architecture used in our framework. For these reason, since our framework has been designed to exploit both body and hands pose information, we recreated all the pose annotations for the NTU RGB+D dataset using the OpenPose network to predict both body and hands joints. For each RGB frame in the NTU RGB+D dataset, we run the OpenPose pose estimator to predict the 2D poses of the people in the images. Then, by using the associated depth frame, the corresponding 3D pose is computed by means of re-projection as described in Section 3.

The authors of the NTU RGB+D dataset proposed two benchmarks: a cross-subject benchmark where all the subjects are split into training and testing groups, and a cross-view benchmark where data from different cameras are used as train and test data. However, in this work, we are interested in using the NTU dataset primarily as a tool to train our models on a large collection of actions; for this reason, we did not follow any of the proposed divisions but tried to use a training set with the largest number of images. We took as training set all the data from the cameras 2 and 3, adding also the data from camera 1 reserved for training in the cross-subject benchmark; the remaining data, namely the test data in the cross-subject benchmark acquired with the camera 1, are used as a small validation set to monitor the training and avoid overfitting.

With such division of the dataset, we trained the models described in Section 3 on the sequences of 3D skeletons derived from OpenPose outputs, considering the subset of the most important joints shown in Figure 2. In particular, we trained a separate model for each set of joints (e.g., *body* joints, *left hand* and *right hand* joints) and a *whole_body* model considering all the available joints (i.e., *body* and *hand* joints). The results obtained for each trained model on the NTU dtaset are reported in Table 2. We evaluated the models in terms of accuracy, considering a *Top1 accuracy* and *Top5 accuracy*. The former represents the percentage of correctly predicted actions in the test set, while the Top5 accuracy is the percentage of actions whose correct prediction falls in the five highest softmax scores estimated by the network.

As reported in Table 2, the best results are obtained with the model trained only on the 15 selected body joints, while the models trained on the hands' joints achieve very low accuracy. These results were somehow expected considering the actions of the NTU dataset, which includes daily actions (e.g., drinking, eating, reading) where the use of hands is very limited. Many of these actions differ mainly in body posture, while the hands represent only marginal information which is not sufficient for a model trained only on hands joints to distinguish such a large set of actions.

Even the model trained using both body and hands joints achieves lower accuracy than the model with only body information. This result seems to suggest that adding hand information is even detrimental to the model, leading it to confuse more actions than using body information alone: considering the 39 input joints, only 15 joints describe the body pose, while more than half represent

Table 2: Experimental results on the NTU RGB+D dataset for each proposed model. Results are provided in terms of accuracy using the skeleton sequences extracted with OpenPose.

Model	#Joints	Top1 %	Top5 %
wholebody	39	59.44	81.34
body	15	**90.40**	**97.68**
hands	24	36.25	67.63
left hand	12	21.12	49.38
right hand	12	31.90	62.68

hand information which does not provide enough information about the actions to be recognized.

Note that the Shift-GCN architecture achieves a Top1 accuracy of 96.5% on the original NTU dataset with body pose annotations, while our "body" model performs slightly worse with a Top1 accuracy of 90.4%. A detailed and direct comparison between the two results is not possible, since we used slightly different train and test sets. However, the drop in performance could be partly due to our new annotations and how partial inputs to the network are handled: if some of the required joints are missing in the input, the entire input skeleton is discarded. We found this happening on several occasions when using the skeletons estimated by OpenPose, leading us to discard whole sequences when too many skeletons were missing.

5.2 Fine-tuning on the IAS-Lab Collaborative HAR Dataset

Using the NTU RGB+D dataset, we were able to train several action recognition models for recognizing a large number of daily activities. Given the large size of the dataset, these models were able to learn several low and mid-level features to be exploited for our action recognition system in collaborative scenarios. We fine-tuned all the models on the IAS-Lab Collaborative HAR dataset described in Section 4, substituting the final layer of 60 nodes with a layer of 18 nodes to match the size of the new set of actions. In order to preserve the low and mid-level features already learnt, during fine-tuning we froze all the layers' weights except the last ones. In particular, denoting with $\ell i, i \in [1, 10]$ the 10 blocks of the Shift-GCN architecture, for each model we froze all the blocks except for the final ones reported in Table 3. We allowed more blocks to be retrained for models that achieved low accuracy on the NTU dataset (e.g., hand models), since in general low accuracy denotes that poor mid- and high-level features have been learned and more layers' weights should be updated.

For fine-tuning the models on the IAS-Lab dataset we followed a cross-subject benchmark, where the first 5 subjects are used for training and the sixth subject is used for validation. All the fine-tuned models have been evaluated in terms of Top1 and Top3 accuracy, reported in Table 3. Since the number of actions in this case is much smaller than in the NTU dataset, we chose to measure a Top3 accuracy instead of the Top5 of the previous case.

Table 3: Experimental results on the IAS-Lab Collaborative HAR dataset for each proposed model. Results are provided in terms of accuracy using the models pre-trained on the NTU RGB+D dataset.

Model	#Joints	valid inputs	Top1 %	Top3 %	learnable blocks
wholebody	39	45%	44.00	54.00	$\ell 7$, $\ell 8$, $\ell 9$, $\ell 10$
body	15	100%	**62.22**	**86.67**	$\ell 9$, $\ell 10$
hands	24	45%	50.00	64.00	$\ell 8$, $\ell 9$, $\ell 10$
left hand	12	75%	59.42	73.91	$\ell 6$, $\ell 7$, $\ell 8$, $\ell 9$, $\ell 10$
right hand	12	70%	59.42	71.01	$\ell 6$, $\ell 7$, $\ell 8$, $\ell 9$, $\ell 10$

Among the results reported in Table 3 the best performances are still obtained with the model trained on the body joints which, however, achieves just 62.22% on Top1 accuracy. Interestingly, we obtain better performance for the models trained on the hands joints compared to the NTU dataset, especially for the models considering each hands separately. In many sequences of the IAS-Lab dataset the subjects' actions were done with one hand, while the other one was in a resting position. These represent ambiguous situations for a model that recognizes actions using information from both hands, reducing its overall accuracy.

However, no model can correctly classify all actions in the test set data, highlighting the complexity of the assigned recognition task. As described in Section 4, the IAS-Lab dataset collects several typical actions of a human-robot collaboration scenario, including both general actions (e.g., *pick*, *place*) and hands gestures (e.g., *confirm*, *stop*). In the former case, the actions to be recognized are large movements involving many body parts, such as walking or hammering. Hand gestures, instead, are commonly used to communicate with the robot and involve only movements of the worker's hands while the rest of the body remains mostly stationary. Both the *body* model and the *single hands* models are very accurate in classifying only one of the two types of actions, but fail on the other one, not having enough information. For example, as shown in Figure 5, the *body* model predicts very accurately actions like *place*, *hammer* or *walk*, but struggles to recognize all the gestures based on hands movements (e.g., *hand to*).

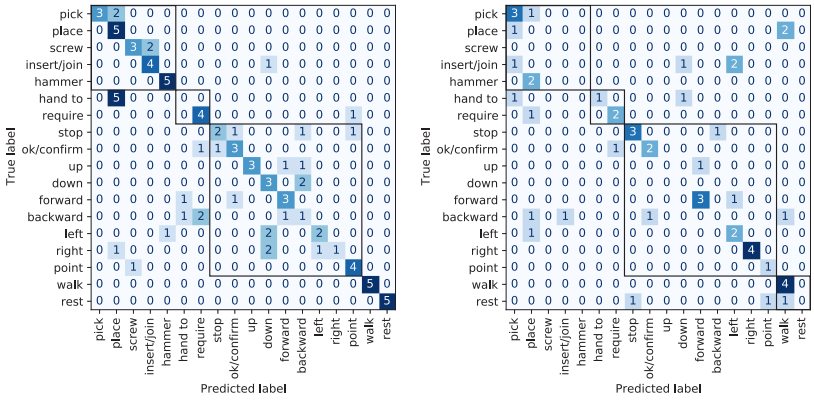

Fig. 5: Confusion matrices for the models fine-tuned on the IAS-Lab Collaborative HAR dataset. On the left, the confusion matrix for the *body* model. On the right, the confusion matrix for the *hands* model.

Intuitively, a model that uses body and hand information together should be the best way to recognize all the given actions, but in our experiments we found the opposite to be true, with the *wholebody* model achieving the lowest accuracy among the results reported in Table 3. This is mainly due to the fact that in several sequences the skeletons found were not complete with all joints, leading to discarding such sequences from the model training set and thus limiting the model's ability to learn to recognize all actions. The percentage of valid sequences available for each models has been reported in Table 3, highlighting how the available data were particularly limited to train the *wholebody* model. Although more accurate in theory, in practice using a model based on the body and hands together proved to be less robust and difficult to train; this observation led us to investigate alternative approaches for combining body and hand information together, such as ensemble techniques.

5.3 Ensemble results using body and hands models

Body posture and hand posture are complementary information that, if properly combined, can greatly aid in the classification of actions. As highlighted in the previous experiment, combining this information when training a network proved to be inefficient: on the one hand, more sequences are required to learn the various relationships that may exist between body and hands during the actions of interest; on the other hand, it requires that in each frame all joints are always detected, which could be not true in case of occlusions.

For these reasons, we developed our action recognition system by combining information at the score level by means of ensemble techniques: we run in parallel the models fine-tuned on the IAS-Lab dataset, and combine together their score predictions (i.e. the predicted probability for each actions) by means of a

weighted sum. In particular, we investigated two main approaches: an ensemble of the *body* and *hands* models, and an ensemble of the *body* model with the models for each hand. In the former case we found that best results are obtained when weighting equally the contributions, that is using weights $\alpha_b = \alpha_h = 0.5$ in Equation (1). When using instead a separate model for each hand we found that less importance should be given to the scores of the *left_hand* model, probably due to the fact that the majority of the subjects in the dataset were right-handed and tended use their rigth hand to perform the requested actions; the final weights we selected for this approach are $\alpha_b = \alpha_{rh} = 0.358$ and $\alpha_{lh} = 0.284$.

Table 4: Experimental results on the IAS-Lab Collaborative HAR dataset using different ensembles of the fine-tuned models.

Model	Top1 %	Top3 %
body + hands	68.00	**90.00**
body + left hand + right hand	**76.54**	87.65

The results obtained using the two ensemble strategies proposed are reported in Table 4. Both strategies lead to an improvement of the Top1 accuracy with respect to the previous results, showing how in general combining body and hands information helps classifying actions. The best result is obtained with the ensemble of the *body* model with the single models for each hand, achieving a huge improvement with respect to the results obtained by each model in Table 3. Although better on the Top3 accuracy, the *body + hands* ensemble achieves a lower Top1 accuracy, mainly due to the many one-handed actions in the dataset, since the second hand not performing the "real" action is misleading for the model. Moreover, such ensemble is also limited by the low number of valid input sequences available, as reported in Table 3. The *body + single hands* ensemble approach represents instead a more robust solution for implementing our action recognition system, since it relies on the base models with the lowest number of joints required: a sequence is considered a valid input if at least the body or a hand is detected, reducing the number of overall frames and sequences discarded.

6 Conclusions

In this work, we propose a unified framework for action recognition in human-robot collaboration scenarios. Our framework can recognize many body-actions and hand-gestures which are often used in several collaborative tasks, thus representing a general solution to various possible real applications. The system has been evaluated on a novel dataset including general actions of human-robot collaboration scenarios, which could be used as a benchmark to further drive

research in this field. In our experiments, we demonstrated how ensemble techniques help to achieve higher accuracy when averaging the predictions of several skeleton-based classifiers, each one trained to recognize actions from a different set of joints. Best results have been obtained using an ensemble of classifiers for body joints and single hands joints, which also represents a robust solution to handle possible missing joints in the estimated skeletons. As future research directions, we will apply the proposed framework in a real human-robot collaboration task to monitor human movements during an assembly process. Moreover, we will investigate the use of a multi-camera setup to improve the 3D pose estimation step, reducing inaccuracies and possible missing joints due to occlusions.

Acknowledgment

The research leading to these results has received funding from the European Unions Horizon 2020 research and innovation program under grant agreement No. 101006732. Part of this work was supported by MIUR (Italian Minister for Education) under the initiative "Departments of Excellence" (Law 232/2016).

References

1. Villani, V., Pini, F., Leali, F., Secchi, C.: Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. Mechatronics **55** (2018) 248–266
2. Matheson, E., Minto, R., Zampieri, E.G., Faccio, M., Rosati, G.: Human–robot collaboration in manufacturing applications: a review. Robotics **8**(4) (2019) 100
3. Kim, W., Peternel, L., Lorenzini, M., Babič, J., Ajoudani, A.: A human-robot collaboration framework for improving ergonomics during dexterous operation of power tools. Robotics and Computer-Integrated Manufacturing **68** (2021) 102084
4. Liu, H., Fang, T., Zhou, T., Wang, L.: Towards robust human-robot collaborative manufacturing: Multimodal fusion. IEEE Access **6** (2018) 74762–74771
5. Mohammadi Amin, F., Rezayati, M., van de Venn, H.W., Karimpour, H.: A mixed-perception approach for safe human–robot collaboration in industrial automation. Sensors **20**(21) (2020) 6347
6. Kobayashi, T., Aoki, Y., Shimizu, S., Kusano, K., Okumura, S.: Fine-grained action recognition in assembly work scenes by drawing attention to the hands. In: 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), IEEE (2019) 440–446
7. Liu, K., Zhu, M., Fu, H., Ma, H., Chua, T.S.: Enhancing anomaly detection in surveillance videos with transfer learning from action recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. (2020) 4664–4668
8. Prati, A., Shan, C., Wang, K.I.K.: Sensors, vision and networks: From video surveillance to activity recognition and health monitoring. Journal of Ambient Intelligence and Smart Environments **11**(1) (2019) 5–22
9. Ranieri, C.M., MacLeod, S., Dragone, M., Vargas, P.A., Romero, R.A.F.: Activity recognition for ambient assisted living with videos, inertial units and ambient sensors. Sensors **21**(3) (2021) 768

10. Al-Amin, M., Tao, W., Doell, D., Lingard, R., Yin, Z., Leu, M.C., Qin, R.: Action recognition in manufacturing assembly using multimodal sensor fusion. Procedia Manufacturing **39** (2019) 158–167
11. Bo, W., Fuqi, M., Rong, J., Peng, L., Xuzhu, D.: Skeleton-based violation action recognition method for safety supervision in the operation field of distribution network based on graph convolutional network. CSEE Journal of Power and Energy Systems (2021)
12. Chen, C., Jafari, R., Kehtarnavaz, N.: Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International conference on image processing (ICIP), IEEE (2015) 168–172
13. Yu, J., Gao, H., Yang, W., Jiang, Y., Chin, W., Kubota, N., Ju, Z.: A discriminative deep model with feature fusion and temporal attention for human action recognition. IEEE Access **8** (2020) 43243–43255
14. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., Baik, S.W.: Action recognition in video sequences using deep bi-directional lstm with cnn features. IEEE access **6** (2017) 1155–1166
15. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 203–213
16. Wen, X., Chen, H., Hong, Q.: Human assembly task recognition in human-robot collaboration based on 3d cnn. In: 2019 IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER), IEEE (2019) 1230–1234
17. Xiong, Q., Zhang, J., Wang, P., Liu, D., Gao, R.X.: Transferable two-stream convolutional neural network for human action recognition. Journal of Manufacturing Systems **56** (2020) 605–614
18. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199 (2014)
19. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1) (2019) 172–186
20. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020)
21. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020) 143–152
22. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021)
23. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1010–1019
24. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C.: Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 2649–2656
25. Martins, G.S., Santos, L., Dias, J.: The growmeup project and the applicability of action recognition techniques. In: Third workshop on recognition and action for scene understanding (REACTS). Ruiz de Aloza. (2015)