International Journal of Neural Systems, Vol. 33, No. 9 (2023) 2350047 (19 pages)

DOI: 10.1142/S0129065723500478

World Scientific
www.worldscientific.com

A Deep Regression Approach for Human Activity Recognition Under Partial Occlusion

Ioannis Vernikos*, Evaggelos Spyrou† and Ioannis-Aris Kostis‡
Department of Informatics and Telecommunications
University of Thessaly, 3rd Km Old National Road Lamia-Athens
Lamia 35132, Greece
**ivernikos@uth.gr*
†*espyrou@uth.gr*
‡*igkostis@teemail.gr*

Eirini Mathe
Department of Informatics, Ionian University
7 Tsirigoti Square, Corfu 49100, Greece
c17math@ionio.gr

Phivos Mylonas§
Department of Informatics and Computer Engineering
University of West Attica, Egaleo Park
Agiou Spyridonos Street
12243 Egaleo, Athens, Greece
mylonasf@uniwa.gr

Accepted 21 June 2023
Published Online 19 August 2023

In real-life scenarios, Human Activity Recognition (HAR) from video data is prone to occlusion of one or more body parts of the human subjects involved. Although it is common sense that the recognition of the majority of activities strongly depends on the motion of some body parts, which when occluded compromise the performance of recognition approaches, this problem is often underestimated in contemporary research works. Currently, training and evaluation is based on datasets that have been shot under laboratory (ideal) conditions, i.e. without any kind of occlusion. In this work, we propose an approach for HAR in the presence of partial occlusion, in cases wherein up to two body parts are involved. We assume that human motion is modeled using a set of 3D skeletal joints and also that occluded body parts remain occluded during the whole duration of the activity. We solve this problem using regression, performed by a novel deep Convolutional Recurrent Neural Network (CRNN). Specifically, given a partially occluded skeleton, we attempt to reconstruct the missing information regarding the motion of its occluded part(s). We evaluate our approach using four publicly available human motion datasets. Our experimental results indicate a significant increase of performance, when compared to baseline approaches, wherein networks that have been trained using only nonoccluded or both occluded and nonoccluded samples are evaluated using occluded samples. To the best of our knowledge, this is the first research work that formulates and copes with the problem of HAR under occlusion as a regression task.

Keywords: Human activity recognition; occlusion; deep learning; regression.

§Corresponding author.

1. Introduction

Human activity recognition (HAR) is the task of recognizing the behavior of humans, based on measurements and/or observations. HAR approaches may rely on several types of sensors that are either worn by the subject or placed within her/his environment while observing her/him or capturing interactions with objects. For example, a human activity such as *walking* may be recognized by capturing changes of her/his appearance using a camera,[1] by analyzing inertial measurements from a smartwatch or band she/he wears,[2–4] or by analyzing her/his interactions with pressure sensors placed on the floor.[5] Yet, wearable/worn sensors demonstrate below average usability and are not preferred by users.[6,7] Also sensors placed within the users' environment either require interventions or are expensive. Thus, many low-cost solutions are based on the use of cameras which capture users' motion within the 3D space. Typical HAR applications include video surveillance, human-computer/robot interaction, augmented reality (AR), ambient assisted environments, health monitoring, intelligent driving, gaming and immersion, animation, etc.[8–10]

Inarguably, HAR from motion data is considered as one of the most challenging computer vision-related problems and may be defined as the recognition of some human behavior within an image or a video sequence using visual data deriving from the human's motion into the 3D space. This behavior consists of a series of consecutive activities ("actions"). An activity may be defined as a distinct type of motion, that is performed by a human, with a relatively short temporal duration.[8] Activities are not instantaneous and typically involve multiple body parts. Special types of activities include interactions, which may involve either a human and an object or two humans and group activities, which involve more than one humans. In the context of this work, we will refer to all the aforementioned types simply as "activities." Also, note that in this work (a) we do not deal with gestures, which are instantaneous and involve at most a couple of body parts; (b) we work with segmented sequences, i.e. each containing exactly one action to be recognized.

HAR approaches based on cameras are typically low-cost and demonstrate more than satisfactory performance in laboratory conditions. However, in real-life situations they suffer from three main problems, which compromise their performance, i.e. viewpoint and illumination changes and occlusion. Viewpoint variance may occur e.g. when the subject is observed by a different viewpoint than the one(s) involved in training. In previous work,[11] we dealt with this problem and demonstrated that the decrease of accuracy due to viewpoint changes may be compensated when using more than one cameras. Illumination changes causing low-light conditions mainly affect video-based approaches. Of course, recent advances in technology have allowed for camera sensors that also capture depth information (which is invariant to illumination changes) and perform significantly better in low-light conditions. In that case, fusion of video and depth data may result in robust extraction of human figures, e.g. as a set of 3D points.[12]

Therefore, from the three aforementioned problems, occlusion is the one that imposes most limitations. In real-life applications, partial or full occlusion may occur e.g. due to furniture or the presence of other humans. Of course, full occlusion renders HAR impossible. However, partial occlusion may significantly affect the accuracy of recognition yet depending on the subset of the visible skeleton, recognition is still possible. In previous work[13] we assessed the effect of partial occlusion in a HAR task and in case a classification model is trained using nonoccluded data and is evaluated using only occluded data. Specifically, we simulated occlusion by removing structured sets of captured moving human skeletons that corresponded to one/two body parts (i.e. arms and legs) and showed that partial occlusion of the subject, in certain cases significantly affected the accuracy of recognition, although the performance drop was tightly depending on each activity.

To tackle the limitations imposed by partial occlusion, in this work, we aim to reconstruct occluded data, upon formulating this problem as a regression task. To this, we use a deep neural network approach, whose input is a human skeleton, with one or more body parts removed, so as to simulate occlusion. We consider the cases of partial occlusion of (a) an arm; (b) a leg; (c) both arms; (d) both legs; (e) an arm and a leg of the same side. We train a convolutional recurrent neural network (CRNN), so as to output the skeleton upon estimating missing part(s).

Its input consists of raw 3D skeleton joint positions, upon removing those joints that correspond to the occluded part(s). Moreover, its output consists of the reconstructed skeleton, which is then fed to a long short term memory (LSTM) network, whose role is to classify it into one of the pre-defined activities. We train a network per occlusion case and we evaluate our approach using four publicly available datasets. To the best of our knowledge, our work is the first that formulates occlusion of moving body parts (i.e. skeleton joint subsets) as a regression task.

The rest of this paper is organized as follows. In Sec. 2, we present research works that deal with the effect of occlusion in HAR-related scenarios. Then, in Sec. 3, we present the proposed regression methodology. Experimental results are presented in Sec. 4. Finally, conclusions are drawn in Sec. 5, wherein plans for future work are also presented.

2. Related Work

During the last few years, a plethora of research works focusing on HAR, based on skeletal data have been presented.[14–20] Moreover, an extensive survey may be found in the work of Wang *et al.*[8] However, although it is widely accepted that occlusion consists one of the most important factors that compromise the performance of HAR approaches,[21] resulting to poor or even unusable results, few are those works that focus either on studying its effects on the performance of recognition or even attempt to overcome them.

To begin with, in the work of Iosifidis *et al.*,[21] a multi-camera setup, surrounding the subject was used for HAR. In order to simulate occlusion, they first trained their algorithm using data from all available cameras and then evaluated using a randomly chosen subset. More specifically, they made the assumption that due to occlusion, not all cameras were simultaneously able to capture the subject's motion. However, we should note that in all cases more than one cameras were able to capture the whole body of the subjects. Also, recognition of a given activity took place upon combining results only from those cameras that were not affected at any means by occlusion. In the work of Gu *et al.*,[22] randomly generated occlusion masks were used in both training and evaluation; each mask caused the occlusion of more than one 2D skeletal joints. Then, and in order to reconstruct the skeleton, they used a regression network. Note that their approach was limited to pose estimation.

Liu *et al.*[23] studied two augmentation strategies for modeling the effect of occlusion. The first discarded independent keypoints, while the second discarded structured sets of keypoints, i.e. those composing main body parts. Note that in this work occluded samples were included in the training process. Moreover, the authors herein made the assumption that the torso and the hips were always visible. Their recognition approach was based on learning view-invariant, occlusion-robust probabilistic embeddings. Similarly, Angelini *et al.*[24] also included artificially occluded samples within the training process. In that case, samples were created by randomly removing body landmarks according to a binary Bernoulli distribution. Their recognition approach was based on pose libraries which included several pose prototypes. When dealing with missing body parts, they exploited the aforementioned libraries either by matching occluded sequences to pre-defined prototypes, based on high-level features, or by filling missing parts upon searching through the pose libraries. In case of short-time occlusions, they used an interpolation approach.

Finally, in previous work[13] we performed a study, wherein our main goal was to assess the effect of occlusion of body parts, within a HAR approach. To this, we created artificial occluded activity samples, by manually removing one or two body parts (i.e. upon removing subsets of skeleton joints). We then extended this work by performing initial experiments using regression on skeletal data.[25]

3. Methodology

3.1. *Skeletal data*

As in previous work,[11,13,25,45] the proposed approach uses as input 3D trajectories of human skeletons. In 3D HAR problems, subjects perform actions in space and over time. We consider skeleton representations as sets of 3D joints. We use skeleton data that have been captured using the Microsoft Kinect RGB/depth camera.[a] A human skeleton comprises of 20 (Kinect v1) or 25 (Kinect v2) joints, organized as a

[a] https://developer.microsoft.com/en-us/windows/kinect.

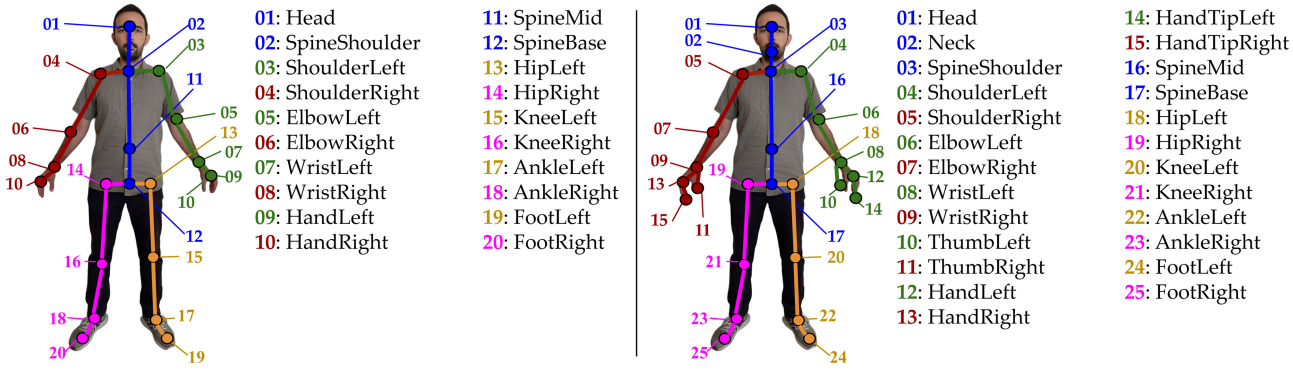


Fig. 1. (Color online) A human body pose with the 20 and 25 skeletal joints that are extracted using the Microsoft Kinect v1 (left) and v2 (right) cameras. Joints have been divided into subsets, each corresponding to one of the five main body parts, i.e. torso (blue), left hand (green), right hand (red), left leg (orange) and right leg (magenta). For illustrative purposes and also to facilitate comparisons between the two different versions, body parts have been colored using the same colors. Numbering follows the Kinect SDK in both cases, therefore there exist several differences between the two versions.

graph; each node corresponds to a body part such as arms, legs, head, neck, etc. while edges follow the body structure, appropriately connecting pairs of joints. In Fig. 1, we illustrate a skeleton extracted using Kinect v1 and v2. Note that joints are shown as being grouped; each group corresponds to a body part, i.e. an arm, a leg or the torso. In the context of this work, an activity is considered to be a temporal sequence of 3D skeleton representations. For the sake of explanation, in Fig. 2, we illustrate an example of a successfully reconstructed skeleton which leads to correct classification and an example of an unsuccessfully reconstructed skeleton, which leads to incorrect classification.

3.2. *Occlusion*

As it has already been mentioned in Sec. 1, partial occlusion may compromise the performance of HAR, in real-life scenarios. Within the context of several applications such as ambient-assisted environments, AR environments, etc. occlusion typically occurs due to e.g. activities taking place behind furniture, or due to the presence of more than one people in the same

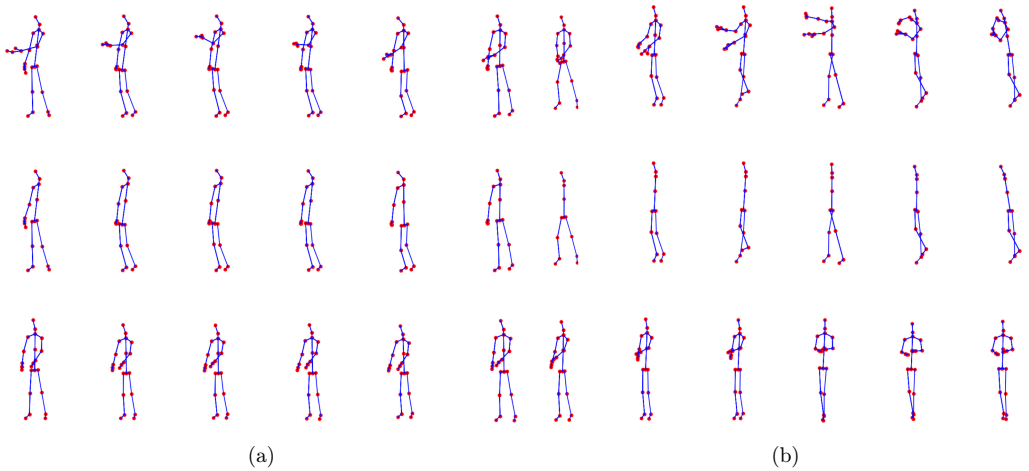


Fig. 2. Example skeleton sequences of the activities (a) *handshaking* and (b) *hugging other person* from the PKU-MMD dataset, captured by Microsoft Kinect v2. First row: original skeletons, including all 25 joints (i.e. without any occlusion); Second row: joints corresponding to (a) left arm; (b) both arms (see Fig. 1) have been discarded (i.e. the skeleton is partially occluded); Third row: skeletons have been reconstructed using the proposed deep regression approach. The example of (a) is successfully reconstructed and correctly classified, while the example of (b) is unsuccessfully reconstructed and incorrectly classified.

room. Of course, it should be obvious that the effects of occlusion vary, depending on the activity performed. For example, the occlusion of both legs when the subject performs the action "kicking" results to a significant loss of visual information, which in turn may result to failure of recognition, while the occlusion of both arms is not expected to compromise recognition for this activity.

Although the aforementioned example is quite extreme, it is common sense that partial occlusion may hinder the overall effectiveness of HAR approaches. We should note that most public motion-based datasets, such as the ones we herein use for evaluation purposes, have been created under ideal laboratory conditions, thus occlusion is prevented. Since the creation of a large scale dataset is a time consuming task, we decided to follow an approach such as the one of Gu *et al.*[22] More specifically, we manually discard subsets of joints that correspond to body parts, assuming that the these parts remain occluded during the whole duration of each activity. The division of a human skeleton into five parts is illustrated in Fig. 1. Moreover and for the sake of explanation, a visual example of an activity upon occlusion is illustrated in Fig. 2, where the loss of visual information is easily comprehensible.

3.3. *Regression of skeletal data*

The input of our approach consists of temporal sequences of 3D skeleton data, i.e. as described in Sec. 3.1. Upon imposing a linear interpolation step between consecutive time frames so as to address temporal variability of activities, we set the length of all activity examples equal to T_m, which denotes the size of the longest one in duration. Note that, if the desired length is not reached upon one interpolation step, the process is repeated until the desired length is reached. As we will mention in Sec. 4, we use datasets that have been captured either using one or three cameras. In the latter case and as we wish to exploit all existing information, we use the corresponding three skeleton sequences as input. We also assume that in every case of occlusion, the same missing body part(s) is (are) occluded in all available cameras.

The core philosophy of our approach is that since occlusion practically causes missing values (i.e. in our case some of joints of the skeleton have been removed), we may formulate the problem of "reconstructing" those missing values as a regression task. More specifically, let $\mathbf{X}$ denote the original skeleton sequence and $\mathbf{X}_o$ the sequence resulting upon occlusion. The goal of regression is ideally to estimate a set of parameters β of a given function f, so that $\mathbf{X} = f(\mathbf{X}_o, \beta) + \epsilon$, where ϵ is some error value, to be minimized.

To this goal, we use a Convolutional Recurrent Neural Network (CRNN) model, whose aim is to implement f and learn β (i.e. its weights), in order to minimize ϵ. Given an occluded skeleton sequence $\mathbf{X}_o$, the network outputs the reconstructed skeleton sequence $\mathbf{X}_r$, which is an estimate for $\mathbf{X}$, upon completing missing (occluded) data (joints). The architecture of the network is as follows: the input of the network constitutes of sequential data from 1 or 3 cameras, depending on the dataset used, as it has already been discussed. We now describe the CRNN architecture for the case of single-camera datasets: in all cases, the duration of each activity is set equal to T_m, each skeleton comprises 20 3D joints (when using Kinect v1), i.e. 60 co-ordinates, in total. Thus, the input layer size is $T_m \times 60$. This is filtered by a stack of 2 2D convolutional layers, followed by a max-pooling layer that performs 1×2 sub-sampling. This single tensor is again filtered by a stack of 2 2D convolutional layers, followed by a max-pooling layer that performs 1×2 sub-sampling. The output of this layer constitutes the input to an LSTM layer, whose goal is to harness temporal information of skeletal data. Then, a dense layer(s) of size $T_m \times 60$ follows and is ultimately reshaped to a $T_m \times 60$ output layer. In case of three-camera datasets, the input of the network constitutes of sequential data from three cameras, each providing a skeletal sequence under a different viewpoint. Those three input branches are in this case each filtered by a stack of 2 2D convolutional layers, followed by a max-pooling layer that performs 1×2 sub-sampling. This process repeats after the three branches are concatenated into a single tensor. This single tensor is again filtered by a stack of 2 2D convolutional layer, followed by a max-pooling layer that performs 1×2 sub-sampling. The output of this layer also in this case constitutes the input to an LSTM layer, and then, three parallel dense layers follow. They are ultimately reshaped to three output layers. The dimension of the kernels of all convolutional layers is 3×3.

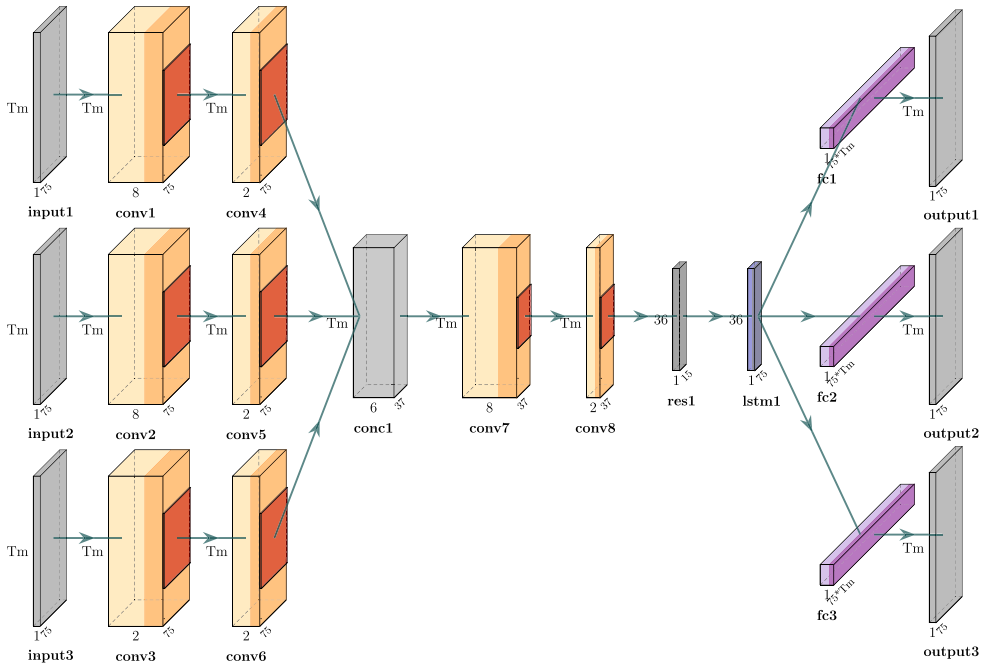


Fig. 3. The architecture of the CRNN, for the three-camera case.

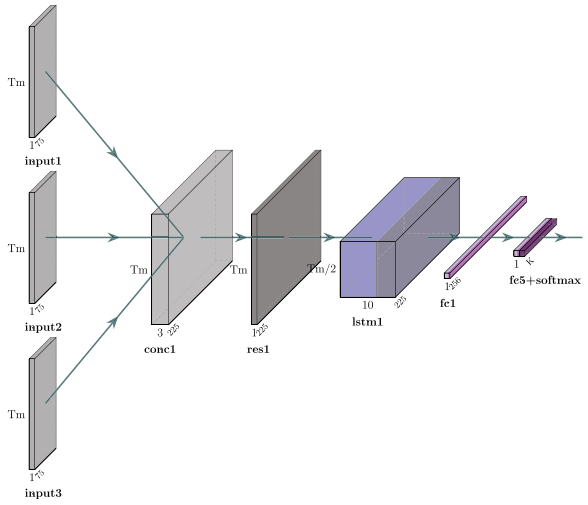


Fig. 4. The architecture of the LSTM network, for the three-camera case.

For loss computation, in both cases, the Mean Square Error (MSE) has been used. In Figs. 3 and 4, we illustrate the architectures of the CRNN and the LSTM networks, respectively, and for the three-camera case.

At this point, we would like to note that the reason for the use of an asymmetrical kernel (i.e. 1×2) is that while being sub-sampled this way, information is compressed only along the spatial coordinates' axes, leaving temporal information intact. We experimentally verified that this kernel choice led to a significant improvement in the performance of the network. In case of three-camera datasets, input from each camera is independently filtered by a stack of 2 2D convolutional layers, followed by a max-pooling layer that performs 1×2 sub-sampling, then these three branches are concatenated into a single tensor and the network's architecture remains the same as in the one-camera case.

The occluded data $\mathbf{X}_o$ are given as input in both training and testing phases of the network. Also, the targets of the network are the nonoccluded data $\mathbf{X}$; these data are to be estimated by the network, i.e. its output are reconstructed data $\mathbf{X}_r$. Thus, the network is trained to learn β, while minimizing ϵ. As we mentioned in Sec. 3.1, in case of using the Kinect SDK in order to extract the skeleton data, each skeleton joint is represented by its own ID. Therefore, in a real-life application, one could easily continuously observe the values of the 3D coordinates and consider those joints with missing/zero values as occluded. This way, the specific case of occlusion could be easily recognized. Bearing this in mind, and upon initial experiments wherein we used a single

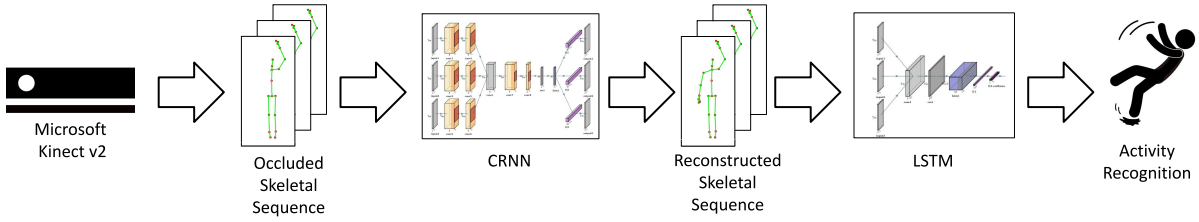


Fig. 5. A visual overview of the proposed approach.

network for all cases, which demonstrated unsatisfactory results, we finally opted to train one network per occlusion case, ending up with eight different networks. Therefore, given an input skeletal sequence, first missing joints are identified and then, it is fed to the appropriate trained network so as to be classified into an activity class, accordingly. Although this approach requires much more memory, its significant increase of performance compared to the use of a single network compensates this requirement.

At this point, an occluded sample (i.e. with missing skeletal joints due to occlusion) may be used as input into the trained CRNN network, which in turn reconstructs the missing skeletal data. For the sake of explanation, a visual example of an activity prior to and upon reconstruction is illustrated in Fig. 2. Upon its reconstruction, we are then able to proceed with the classification of the aforementioned into one of the pre-defined classes. For classification we use an LSTM network with one layer. Note that the CRNN network is trained using only the non-occluded data $\mathbf{X}$, thus without with any occlusion information. On the contrary, during the testing phase, the input of the LSTM network are the corresponding skeletal sequences $\mathbf{X}_r$, which have been reconstructed by $\mathbf{X}$. A visual overview of the proposed approach is illustrated in Fig. 5.

4. Experiments and Results

4.1. *Datasets*

To the best of our knowledge, there does not exist any publicly available dataset comprising 3D occluded actions. To overcome this, we manually discarded subsets of joints that correspond to body parts (i.e. arms and legs) from four publicly available datasets. Specifically, we have used the following:

PKU-MMD[26] is a public and open benchmark dataset focusing on 3D human motion-based activity understanding, recorded under three camera viewpoints. In this work, we selected a set of 11 classes and 21,456 samples, which correspond to activities of daily living,[1,27] i.e.: *eat meal snack* (10), *falling* (11), *handshaking* (14), *hugging other person* (16), *make a phone call answer phone* (20), *playing with phone tablet* (23), *reading* (30), *sitting down* (33), *standing up* (34), *typing on a keyboard* (46) and *wearing a jacket* (48).

NTU-RGB+D[28] is a large-scale benchmark dataset for 3D human activity analysis, also recorded under three camera viewpoints. We selected the medical-condition-related category consisting of 12 classes and 11,400 samples, namely: *sneeze/cough* (41), *staggering* (42), *falling down* (43), *headache* (44), *chest pain* (45), *back pain* (46), *neck pain* (47), *nausea/vomiting* (48), *fan self* (49), *yawn* (101), *stretch oneself* (102) and *blow nose* (103).

SYSU 3D Human-Object Interaction (HOI)[29] is a small-scale dataset focusing on 3D human motion-based interactions between humans and objects, comprising 480 activity instances from 12 different activities that involve interactions of 40 subjects with one of the following objects: phone, chair, bag, wallet, mop and besom; specifically 40 instances from each of the following activities are provided: *drinking* (1), *pouring* (2), *calling phone* (3), *playing phone* (4), *wearing backpacks* (5), *packing backpacks* (6), *sitting chair* (7), *moving chair* (8), *taking out wallet* (9), *taking from wallet* (10), *mopping* (11) and *sweeping* (12).

UTKinect-Action3D dataset[30] consists of 10 simple activities that have been performed by 10 subjects; each performing all activities twice, i.e. 200 activity instances are provided from the following activities *walk* (0), *sitDown* (1), *standUp* (2), *pickUp* (3), *carry* (4), *throw* (5), *push* (6), *pull* (7), *waveHands* (8), *clapHands* (9); each category comprises 20 instances.

Note that in all cases, numbers in parentheses denote the corresponding class ids, following the

original notation per case and will be used at the remaining of this paper to facilitate readability when necessary. Also, note that in all cases we relied only on 3D skeleton data and discarded other modalities. PKU-MMD and NTU-RGB datasets have been captured using Microsoft Kinect v2, while SYSU-3D-HOI and UTKinect-Action3D have been captured using Microsoft Kinect v1.

4.2. *Experimental setup and network training*

Experiments were performed on a personal workstation with an Intel™ i7 4770 4-core processor on 3.40 GHz and 16 GB RAM, using NVIDIA™ Geforce RTX 2060 Super GPU with 8 GB VRAM and Ubuntu 20.04 (64 bit). The deep architecture has been implemented in Python, using Keras 2.4.3[31] with the Tensorflow 2.5[32] backend. All data preprocessing and processing steps have been implemented in Python 3.9 using NumPy and SciPy. For the training of the estimator, we used the LeakyReLU activation function, except from the LSTM layer wherein the tanh function was used, and the last dense layer wherein linear activation function was used. For the training of the classifier, the LeakyReLU and tanh activation functions were used, respectively, except from the last layer, wherein the softmax activation function was used. Moreover, we set the batch size to 5 and 10 for the training of the classifier and the estimator, respectively. The Adam optimizer was utilized in both cases, the dropout was set to 0.3, the learning rate to 0.001 and we trained for 50 epochs, using the loss of the validation set calculated via MSE as an early stopping method, in order to prevent overfitting. In all cases, we used 80% of the available data for training, 10% for validation and the remaining 10% for testing. The code for replicating our experiments is publicly available on GitHub.[b]

4.3. *Evaluation protocol and results*

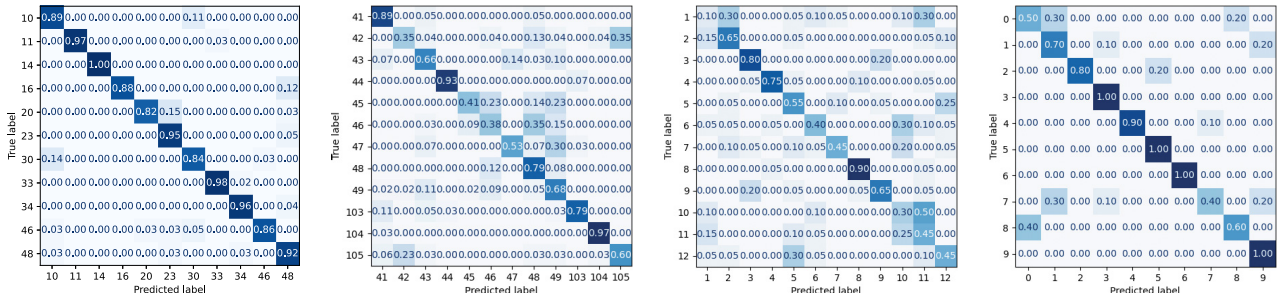
For the experimental evaluation of the proposed methodology, we considered the following:

(a) Removal of structured sets of skeletal joints, corresponding to body parts, to simulate occlusion (see Fig. 1). Specifically, we removed (a) left arm; (b) right arm; (c) both arms; (d) left leg; (e) right leg; (f) both legs; (g) left arm and left leg; (h) right arm and right leg. In this experiment we evaluate the performance of classification of reconstructed samples, using an LSTM network that has been trained using nonoccluded samples, while reconstruction takes place using the CRNN,
(b) A baseline approach, wherein the LSTM is trained and evaluated using nonoccluded samples,
(c) A "reference" approach wherein the LSTM is trained using nonoccluded samples and validated using occluded samples, and
(d) Inclusion of occluded samples in the training process of the LSTM and validation using occluded samples. In that case, a subset equal to 10% of the nonoccluded samples of training set is selected. From those samples we create all eight cases of occlusion, therefore we "augment" the initial training set by 80%. Note that in this case and contrary to the first of the aforementioned cases, we used a single network for all eight cases of occlusion.

Experimental results for all datasets are depicted in Tables 1–4. In all cases we extract the following metrics: accuracy per class, F_1 score per class and weighted accuracy, where class weights are calculated based on the class distribution. Moreover, confusion matrices for all datasets in case of the baseline experiment are depicted in Fig. 6, while in case of classification of reconstructed samples are depicted in Figs. 7–10.

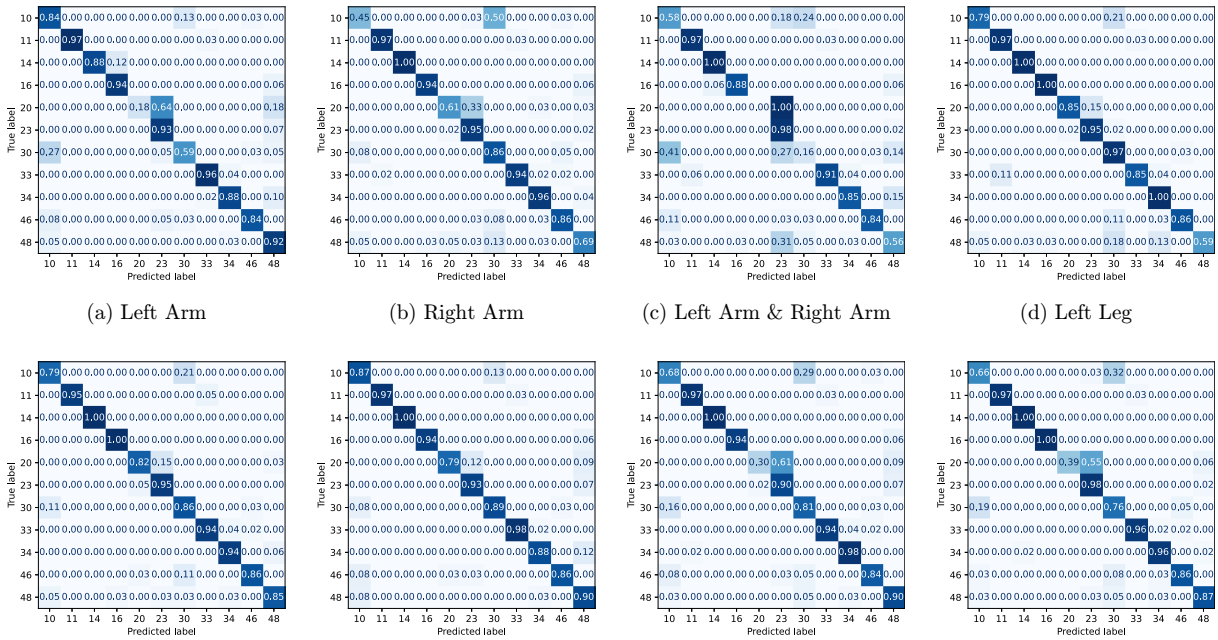
In case of the PKU-MMD dataset, the weighted accuracy (WA) was 0.92 without any body part removal. Specifically, it ranged between 0.21–0.90 in case of some body part removal, while it ranged between 0.70–0.91 upon reconstruction. In 7 out of 8 cases, significant improvement was observed, in terms of WA, while performance was almost equal in case of removal of left leg. Moreover, reconstruction outperformed occlusion in 6 out of 8 cases. Intuitively, one should observe that the majority of the activities we used to evaluate our approach mainly consists of upper body motion (i.e. left and/or right arm). Upon careful observation of the samples of the datasets, this assumption has been verified. This is also reflected to the results of Table 1, wherein it may observed that in cases of occluded arms the improvement is significantly large, with most notable

[b] https://github.com/thevisionlab-uth/HAR_Regression.



(a) PKU-MMD (b) NTU-RGB+D (c) SYSU-3D-HOI (d) UTKinect-Action-3D

Fig. 6. Normalized confusion matrices for classification for all datasets, without removing any body part.



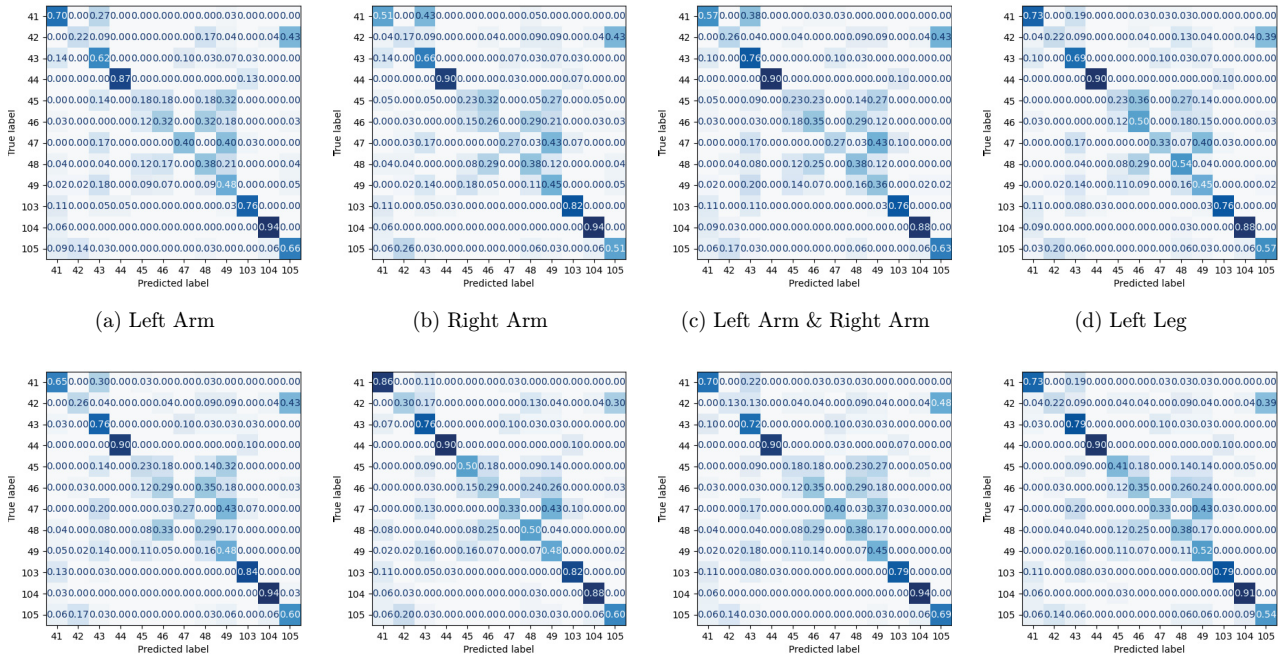
(a) Left Arm (b) Right Arm (c) Left Arm & Right Arm (d) Left Leg

(e) Right Leg (f) Left Leg & Right Leg (g) Left Arm & Left Leg (h) Right Arm & Right Leg

Fig. 7. Normalized confusion matrices for classification for the PKU-MMD dataset, upon removing the body part(s) denoted in the caption of the corresponding subfigure.

example the case of both arms, wherein WA improves from 0.21 to 0.70. Upon careful observation of the confusion matrices depicted in Fig. 7, for each occlusion case we should notice the following, when comparing with the case where all joints had been used: (a) in case of any occluded arm, class *make a phone call/answer phone* is often confused with *playing with phone/tablet* and class *eat meal/snack* is often confused with *reading*; (b) in case of occluded left leg, class *wear jacket* is often confused with *reading* or *standing up*; and (c) finally, in case of both arms occluded, 7 classes show adequate performance. Upon observing the performance of individual classes, we should highlight the following: in the baseline experiment, all classes are recognized with high accuracies/F_1 scores. Exception in this case is the class *make a phone call answer phone*, in case of occluded arm(s) which demonstrates extremely small values of both these metrics, as expected.

In case of NTU-RGB+D dataset, the WA was 0.68 without any body part removal and ranged between 0.16–0.59 in case of some body part removal, while it ranged between 0.53–0.62 upon reconstruction. Also in that case, in 7 out of 8 cases, significant improvement was observed, in terms of WA, while performance was almost equal in case of



(a) Left Arm (b) Right Arm (c) Left Arm & Right Arm (d) Left Leg

(e) Right Leg (f) Left Leg & Right Leg (g) Left Arm & Left Leg (h) Right Arm & Right Leg

Fig. 8. Normalized confusion matrices for classification for the NTU-RGB+D dataset upon removing the body part(s) denoted in the caption of the corresponding subfigure.

removal of left arm. Moreover, reconstruction outperformed occlusion in 7 out of 8 cases. Since activities used to evaluate the proposed approach mainly consisted of upper body motion, also in the case of the NTU-RGB+D dataset, in the results of Table 2 it could be observed that in all the remaining cases of occluded arms, the improvement of WA is large, with the most notable example the case of both arms, wherein WA improves from 0.21 to 0.70. Upon careful observation of the confusion matrices depicted in Fig. 7 and considering the classification results without occlusion of any part, for each occlusion case we should notice that occlusion mainly affects activities *back pain*, *chest pain*, *neck pain*, *nausea/vomiting* and *fan self*. Note that these activities are performed in a significantly different way than the remaining ones and also that those of PKU-MMD, i.e. joints involved perform less movements. However, 7 classes in presence of occlusion consistently exhibit comparable performance to the case of absence of occlusion. Upon observing the performance of individual classes, we should highlight the following: in the baseline experiment, almost all classes are recognized with high accuracies/F_1 scores. Exceptions in this case are the classes *staggering*, *chest pain* and *back pain*, in case of occluded arm(s) which demonstrate extremely small values of both these metrics, as expected.

In case of SYSU-3D-HOI dataset, the WA was 0.54 without any body part removal and ranged between 0.10–0.20 in case of some body part removal, while it ranged between 0.39–0.48 upon reconstruction. Moreover, reconstruction outperformed occlusion in 6 out of 8 cases. In every case of occlusion, significant improvement was observed, in terms of WA. We should herein notice that due to the small size of this dataset, performance was inadequate in several classes even without occlusion. As expected, worst performance was observed in case of occlusion of both arms. However, in the majority of cases, most classes in presence of occlusion consistently exhibit comparable performance to the case of absence of occlusion. Upon careful observation of the confusion matrices depicted in Fig. 9 and considering the classification results without occlusion of any part, for each occlusion case we should notice that occlusion mainly affects activities *pouring*, *calling phone*, *packing backpacks* and *taking from wallet*, which are based on the motion of arms.

Table 1. Results on PKU-MMD dataset.

		None	LA			RA			LA+RA			LL			RL			LL+RL			LA+LL			RA+RL		
Class	Metric	Bas.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.
10	Acc.	0.90	**0.84**	0.43	0.11	**0.45**	0.36	0.03	0.58	**0.64**	0.00	0.79	0.71	**0.92**	**0.79**	0.50	0.66	**0.87**	0.71	0.74	0.68	**0.79**	0.21	0.66	**0.71**	0.24
	F_1	0.86	**0.75**	0.60	0.15	**0.57**	0.32	0.05	0.55	**0.78**	0.00	**0.86**	0.83	0.84	**0.81**	0.67	0.72	**0.82**	0.74	0.78	0.70	**0.88**	0.25	0.69	**0.77**	0.36
11	Acc.	0.97	0.97	**1.00**	0.97	**0.97**	0.85	0.97	0.97	**1.00**	0.00	**0.97**	0.92	0.97	0.95	**1.00**	0.97	**0.97**	0.85	0.97	0.97	**1.00**	**1.00**	0.97	**1.00**	0.97
	F_1	0.99	**0.99**	0.96	0.90	**0.97**	0.88	0.99	0.95	**1.00**	0.00	0.91	0.92	**0.99**	0.97	**1.00**	0.99	**0.99**	0.88	0.99	**0.97**	0.96	0.70	0.99	**1.00**	0.92
14	Acc.	1.00	0.88	1.00	**1.00**	**1.00**	0.29	1.00	**1.00**	1.00	0.81	**1.00**	**1.00**	0.94	**1.00**	0.86	1.00	**1.00**	0.57	1.00	**1.00**	**1.00**	0.94	**1.00**	0.29	1.00
	F_1	1.00	0.93	0.58	**0.97**	**1.00**	0.27	1.00	**0.97**	0.67	0.90	**0.97**	0.64	0.94	**1.00**	0.71	0.91	**1.00**	0.62	0.91	**1.00**	0.78	0.91	**0.97**	0.33	0.86
16	Acc.	0.88	**0.94**	0.14	0.88	**0.94**	0.14	0.88	**0.88**	0.43	0.69	**1.00**	0.29	0.88	**1.00**	0.57	0.88	**0.94**	0.57	0.88	**0.94**	0.57	0.81	**1.00**	0.43	0.88
	F_1	0.93	**0.91**	0.25	0.90	**0.94**	0.25	0.93	**0.90**	0.50	0.76	**0.97**	0.44	0.93	**0.97**	0.73	0.90	**0.97**	0.67	0.90	**0.97**	0.62	0.79	**1.00**	0.50	0.85
20	Acc.	0.82	0.18	**0.91**	0.03	**0.61**	0.73	0.39	0.00	0.73	0.00	0.85	**1.00**	0.79	0.82	0.73	**0.88**	0.79	0.55	**0.88**	0.30	**0.91**	0.00	0.39	0.64	**0.97**
	F_1	0.89	0.31	**0.57**	0.06	**0.71**	0.47	0.19	0.00	0.59	0.00	**0.90**	0.61	0.87	0.86	0.67	**0.62**	0.87	0.55	**0.90**	0.45	**0.71**	0.00	**0.57**	0.45	0.27
23	Acc.	0.95	0.93	0.24	**0.98**	**0.95**	0.35	0.83	**0.98**	0.59	0.05	**0.95**	0.18	0.93	**0.95**	0.82	0.02	0.93	**0.94**	0.05	0.91	0.59	**1.00**	**0.98**	0.53	0.05
	F_1	0.90	**0.74**	0.38	0.70	**0.83**	0.44	0.40	**0.56**	0.69	0.05	0.92	0.30	**0.93**	**0.91**	0.85	0.05	**0.91**	0.84	0.09	**0.75**	0.71	0.62	**0.80**	0.56	0.07
30	Acc.	0.84	0.60	0.86	**0.70**	**0.87**	0.21	0.24	0.16	**0.93**	0.00	**0.97**	0.86	0.76	0.87	**0.93**	0.89	**0.89**	0.79	0.84	0.81	**0.93**	0.30	0.76	**0.86**	0.70
	F_1	0.84	**0.68**	0.71	0.48	**0.67**	0.32	0.36	0.22	**0.79**	0.00	0.77	0.80	**0.78**	**0.78**	0.74	0.75	**0.88**	0.73	0.75	0.73	**0.84**	0.27	0.68	**0.77**	0.68
33	Acc.	0.98	0.96	**1.00**	0.74	**0.94**	0.95	0.06	0.91	**1.00**	0.00	0.85	0.95	**0.98**	0.94	**1.00**	0.98	0.98	**1.00**	0.98	0.94	**0.95**	0.11	0.96	**1.00**	0.00
	F_1	0.98	0.96	**1.00**	0.84	**0.96**	0.92	0.11	0.94	**1.00**	0.00	0.91	0.95	**0.98**	0.95	**1.00**	0.98	**0.98**	0.97	0.98	0.96	**0.97**	0.20	0.97	**1.00**	0.00
34	Acc.	0.96	0.89	0.95	**0.96**	**0.96**	0.95	0.19	0.46	**1.00**	0.00	**1.00**	0.95	0.94	0.94	**1.00**	0.90	0.89	**1.00**	0.85	0.98	**1.00**	0.54	0.96	**1.00**	0.00
	F_1	0.96	0.91	**0.97**	0.96	0.94	**0.97**	0.32	0.89	**1.00**	0.00	0.93	**0.97**	0.95	0.94	**1.00**	0.93	0.92	**1.00**	0.91	0.96	**1.00**	0.70	0.95	**1.00**	0.00
46	Acc.	0.87	0.84	**1.00**	0.89	0.87	**1.00**	0.49	0.84	**1.00**	0.65	0.87	**1.00**	0.84	0.87	**1.00**	0.87	0.87	**1.00**	0.87	0.84	**1.00**	0.87	0.87	**1.00**	0.65
	F_1	0.91	0.89	**0.97**	0.69	**0.88**	0.79	0.38	0.90	**1.00**	0.14	0.91	**0.94**	0.90	0.90	**0.97**	0.93	0.91	**0.97**	0.91	0.87	**0.97**	0.48	0.89	**1.00**	0.56
48	Acc.	0.92	**0.92**	0.80	0.28	0.69	**0.73**	0.13	0.56	**0.67**	0.00	0.59	0.80	**0.85**	0.85	**0.87**	0.85	**0.90**	0.80	0.90	**0.90**	0.73	0.05	**0.87**	0.73	0.13
	F_1	0.88	**0.78**	0.89	0.43	0.76	**0.76**	0.22	0.58	**0.80**	0.00	0.74	0.86	**0.87**	0.87	**0.93**	0.72	**0.80**	0.89	0.74	**0.86**	0.85	0.10	**0.88**	0.85	0.23
all	WA	0.92	**0.82**	0.78	0.68	**0.84**	0.65	0.40	0.70	**0.83**	0.21	0.89	0.80	**0.90**	**0.90**	0.87	0.80	**0.91**	0.84	0.80	0.84	**0.87**	0.48	**0.86**	0.80	0.41

Notes: "Bas." / "Rec."/"Occ."/"Ref." denote baseline / reconstructed / training with occluded samples / reference case, Acc, F_1 and WA denote Accuracy, F_1 score and weighted accuracy, respectively. "None" denotes the case without occlusion. LA, RA, LL, RL denote the occlusion of left arm, right arm, left leg, right leg, respectively. Numbers in bold indicate best performance between "Rec."/"Occ."/"Ref.".

Table 2. Results on NTU-RGB+D dataset.

Class	Metric	None Bas.	LA Rec.	LA Occ.	LA Ref.	RA Rec.	RA Occ.	RA Ref.	LA+RA Rec.	LA+RA Occ.	LA+RA Ref.	LL Rec.	LL Occ.	LL Ref.	RL Rec.	RL Occ.	RL Ref.	LL+RL Rec.	LL+RL Occ.	LL+RL Ref.	LA+LL Rec.	LA+LL Occ.	LA+LL Ref.	RA+RL Rec.	RA+RL Occ.	RA+RL Ref.
41	Acc.	0.89	0.70	**0.87**	0.73	0.51	**0.74**	0.14	0.57	**0.84**	0.11	0.73	**0.84**	0.24	0.65	**0.81**	0.62	**0.87**	0.74	0.24	0.70	**0.84**	0.16	**0.73**	0.71	0.11
	F_1	0.82	0.66	0.73	**0.74**	0.54	**0.65**	0.19	0.58	**0.79**	0.11	0.70	**0.71**	0.35	0.66	**0.74**	0.59	**0.78**	0.73	0.38	0.68	**0.80**	0.24	**0.73**	0.73	0.19
42	Acc.	0.35	0.22	**0.47**	0.13	0.17	0.34	**0.26**	**0.26**	0.41	0.00	0.22	**0.41**	0.09	0.26	**0.44**	0.09	**0.30**	0.44	0.00	0.13	**0.41**	0.00	0.22	**0.47**	0.09
	F_1	0.40	0.29	**0.42**	0.17	0.20	0.33	**0.31**	**0.32**	0.39	0.00	0.28	**0.40**	0.14	0.32	**0.40**	0.15	**0.36**	0.41	0.00	0.19	**0.39**	0.00	0.28	**0.45**	0.14
43	Acc.	0.66	0.62	0.59	**0.72**	0.66	0.53	**0.83**	**0.76**	0.69	0.62	**0.69**	0.47	0.52	**0.76**	0.69	0.76	**0.76**	0.69	0.72	**0.72**	0.69	0.62	**0.79**	0.69	0.55
	F_1	0.61	0.46	0.50	**0.46**	**0.47**	0.42	0.35	**0.49**	0.51	0.29	**0.53**	0.41	0.38	**0.54**	0.46	0.27	**0.57**	0.50	0.36	**0.51**	0.59	0.31	**0.56**	0.51	0.26
44	Acc.	0.93	**0.87**	0.87	0.77	**0.90**	0.87	0.07	**0.90**	0.87	0.07	**0.90**	0.87	0.93	0.90	0.90	**0.97**	0.90	0.84	**1.00**	0.90	0.87	**0.97**	**0.90**	0.84	0.37
	F_1	0.95	**0.90**	0.89	0.87	**0.93**	0.89	0.12	**0.95**	0.89	0.12	**0.93**	0.87	0.93	0.95	0.90	**0.95**	0.93	0.88	**0.85**	**0.93**	0.87	0.89	**0.93**	0.88	0.54
45	Acc.	0.41	0.18	0.28	**0.46**	0.23	0.19	**0.91**	0.23	0.28	**0.41**	0.23	**0.25**	0.14	0.23	0.34	**0.59**	0.50	0.22	**0.50**	**0.18**	0.41	0.14	0.41	0.16	**0.77**
	F_1	0.51	0.22	0.27	**0.40**	0.24	0.23	**0.26**	0.23	0.25	**0.23**	**0.26**	0.25	0.17	0.26	0.29	**0.32**	**0.47**	0.26	**0.23**	**0.21**	0.38	0.19	0.40	0.19	**0.23**
46	Acc.	0.38	0.32	**0.47**	0.24	0.27	**0.47**	0.12	0.35	**0.56**	0.27	0.50	**0.53**	0.32	0.29	**0.44**	0.15	**0.29**	0.50	0.09	**0.35**	0.59	0.12	0.35	**0.66**	0.03
	F_1	0.43	**0.39**	0.35	0.31	0.30	**0.33**	0.17	0.39	**0.41**	0.31	**0.47**	0.36	0.37	0.33	**0.38**	0.21	0.35	**0.36**	0.14	**0.37**	0.41	0.20	0.40	**0.42**	0.05
47	Acc.	0.53	0.40	0.42	**0.63**	0.27	**0.42**	0.27	0.27	**0.58**	0.07	0.33	**0.45**	0.33	0.27	0.52	**0.37**	0.33	**0.65**	0.13	0.40	**0.71**	0.33	0.33	**0.65**	0.43
	F_1	0.64	0.53	0.55	**0.73**	0.39	**0.55**	0.28	0.38	**0.65**	0.11	0.45	**0.56**	0.43	0.39	0.58	**0.48**	0.45	**0.63**	0.21	0.51	**0.70**	0.43	0.45	**0.63**	0.36
48	Acc.	0.79	0.38	0.10	**0.42**	0.38	0.07	0.17	0.38	0.03	**0.50**	0.54	0.07	**0.67**	**0.29**	0.10	0.21	**0.50**	0.10	0.21	0.38	0.07	**0.63**	**0.38**	0.03	0.21
	F_1	0.55	0.31	0.15	**0.47**	0.32	0.10	0.18	**0.31**	0.06	0.14	**0.39**	0.11	0.31	**0.24**	0.16	0.22	**0.44**	0.17	0.14	0.32	0.12	**0.32**	**0.32**	0.06	0.15
49	Acc.	0.68	0.48	0.44	**0.48**	**0.46**	0.38	0.05	0.36	**0.50**	0.00	0.46	0.28	**0.77**	0.48	**0.56**	0.16	**0.48**	0.44	0.41	0.46	0.38	**0.59**	**0.52**	0.41	0.23
	F_1	0.59	0.42	0.44	**0.47**	**0.41**	0.39	0.07	0.36	**0.42**	0.00	0.44	0.35	**0.46**	0.42	**0.47**	0.24	**0.45**	0.38	0.31	0.43	0.40	**0.40**	**0.46**	0.37	0.26
103	Acc.	0.79	0.76	0.72	**0.79**	**0.82**	0.78	0.40	**0.76**	0.53	0.16	**0.76**	0.75	0.74	**0.84**	0.50	0.74	**0.82**	0.69	0.61	**0.79**	0.72	0.74	**0.79**	0.72	0.32
	F_1	0.85	0.79	0.81	**0.79**	**0.84**	0.83	0.54	**0.79**	0.68	0.26	0.82	0.81	**0.84**	0.84	0.67	**0.85**	**0.83**	0.77	0.72	**0.85**	0.79	0.82	**0.83**	0.81	0.48
104	Acc.	0.97	**0.94**	0.75	0.74	**0.94**	0.75	0.21	**0.88**	0.69	0.00	**0.88**	0.84	0.21	**0.94**	0.63	0.15	**0.88**	0.78	0.03	**0.94**	0.78	0.12	**0.91**	0.81	0.32
	F_1	0.96	**0.93**	0.81	0.82	**0.90**	0.76	0.34	**0.88**	0.79	0.00	**0.90**	0.84	0.33	**0.93**	0.75	0.24	**0.90**	0.86	0.06	**0.91**	0.83	0.21	**0.89**	0.85	0.48
105	Acc.	0.60	0.66	0.34	**0.77**	**0.51**	0.38	0.11	**0.63**	0.22	0.00	**0.57**	0.47	0.14	**0.60**	0.16	0.11	**0.60**	0.25	0.00	**0.69**	0.28	0.20	**0.54**	0.28	0.03
	F_1	0.66	0.64	0.37	**0.65**	**0.54**	0.38	0.17	**0.65**	0.27	0.00	**0.61**	0.46	0.24	**0.62**	0.22	0.19	**0.65**	0.31	0.00	**0.69**	0.32	0.26	**0.60**	0.35	0.05
all	WA	0.68	0.57	0.53	**0.59**	**0.53**	0.49	0.27	**0.55**	0.52	0.16	**0.59**	0.52	0.44	**0.57**	0.51	0.41	**0.62**	0.53	0.33	**0.58**	0.56	0.40	**0.59**	0.53	0.25

Notes: "Bas." / "Rec."/"Occ."/"Ref." denote baseline / reconstructed / training with occluded samples / reference case, Acc, F_1 and WA denote Accuracy, F_1 score and weighted accuracy, respectively. "None" denotes the case without occlusion. LA, RA, LL, RL denote the occlusion of left arm, right arm, left leg, right leg, respectively. Numbers in bold indicate best performance between "Rec."/"Occ."/"Ref.".

Table 3. Results on SYSU-3D-HOI dataset.

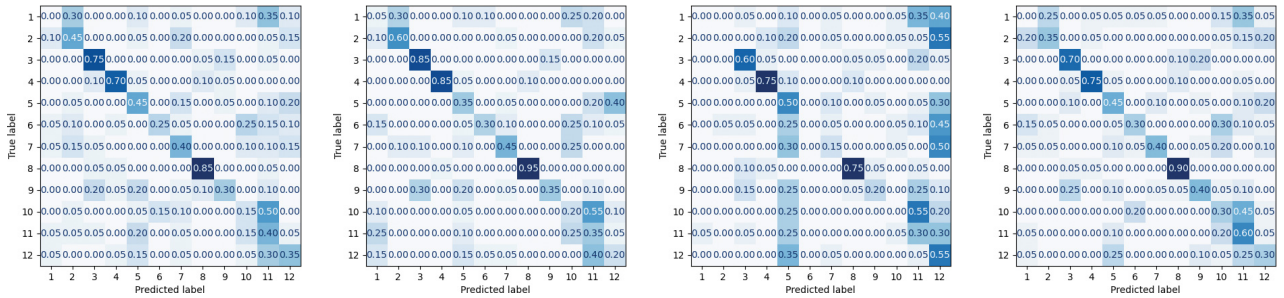
Class	Metric	None Baseline	LA			RA			LA+RA			LL			RL			LL+RL			LA+LL			RA+RL		
			Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.
1	Acc.	0.10	0.00	**0.13**	0.00	0.05	**0.13**	0.00	0.00	**0.25**	0.00	0.00	**0.10**	0.00	0.10	**0.28**	0.00	0.05	**0.13**	0.00	0.00	**0.28**	0.00	0.00	**0.05**	0.00
	F_1	0.10	0.00	**0.14**	0.00	0.06	**0.07**	0.00	0.00	**0.25**	0.00	0.00	**0.11**	0.00	0.10	**0.28**	0.00	0.08	**0.11**	0.00	0.00	**0.28**	0.00	0.00	**0.06**	0.00
2	Acc.	0.65	**0.45**	0.43	0.00	**0.65**	0.13	0.05	0.00	**0.50**	0.00	**0.45**	0.40	**0.45**	**0.55**	0.53	0.20	**0.55**	0.28	0.10	**0.50**	0.48	0.30	0.00	**0.28**	0.00
	F_1	0.59	0.36	**0.46**	0.00	**0.63**	0.12	0.08	0.00	**0.54**	0.00	0.44	**0.45**	0.34	0.56	**0.57**	0.23	**0.53**	0.34	0.04	**0.48**	0.49	0.16	0.00	**0.19**	0.00
3	Acc.	0.80	**0.75**	0.60	0.55	**0.85**	0.48	0.05	0.60	**0.68**	0.45	**0.70**	0.58	0.00	**0.85**	0.68	0.00	**0.80**	0.48	0.00	0.60	**0.65**	0.25	**0.55**	0.55	0.10
	F_1	0.77	**0.69**	0.54	0.21	**0.77**	0.44	0.05	0.57	**0.59**	0.12	**0.65**	0.56	0.00	**0.71**	0.59	0.00	**0.73**	0.46	0.00	**0.56**	0.59	0.18	**0.61**	0.50	0.06
4	Acc.	0.75	**0.70**	0.58	0.05	**0.80**	0.53	0.30	**0.75**	0.65	0.00	**0.75**	0.68	0.40	**0.75**	0.65	0.25	**0.70**	0.58	0.20	**0.75**	0.70	0.25	**0.80**	0.68	0.25
	F_1	0.80	**0.73**	0.61	0.08	**0.84**	0.57	0.43	**0.76**	0.71	0.00	**0.79**	0.70	0.52	**0.78**	0.72	0.34	**0.75**	0.62	0.29	**0.75**	0.76	0.27	**0.76**	0.66	0.31
5	Acc.	0.55	**0.45**	0.28	0.10	0.40	**0.53**	0.40	**0.50**	0.40	0.00	**0.45**	0.38	0.40	**0.40**	0.40	0.00	**0.50**	0.25	0.20	0.30	0.33	**0.50**	**0.50**	0.30	0.10
	F_1	0.48	**0.34**	0.21	0.10	0.35	**0.40**	0.15	**0.30**	0.38	0.00	**0.38**	0.32	0.10	0.39	**0.41**	0.00	**0.39**	0.23	0.19	**0.28**	0.32	0.20	**0.44**	0.31	0.07
6	Acc.	0.40	0.25	**0.28**	0.05	**0.25**	0.05	0.00	0.00	**0.30**	0.00	0.30	**0.33**	0.00	**0.35**	0.28	0.00	**0.25**	0.25	0.00	**0.30**	0.33	0.00	0.25	**0.43**	0.00
	F_1	0.38	**0.31**	0.27	0.08	**0.28**	0.03	0.00	0.00	**0.28**	0.00	**0.27**	0.31	0.00	**0.32**	0.26	0.00	0.18	**0.22**	0.00	**0.27**	0.30	0.00	0.20	**0.38**	0.00
7	Acc.	0.45	0.40	0.43	**0.50**	**0.35**	0.33	0.30	0.15	**0.65**	0.00	0.40	**0.43**	0.00	**0.45**	0.68	0.00	**0.45**	0.45	0.00	0.35	**0.55**	0.20	0.35	**0.53**	0.00
	F_1	0.54	0.35	**0.39**	0.18	**0.36**	0.29	0.19	0.20	**0.61**	0.00	**0.47**	0.42	0.00	**0.46**	0.60	0.00	**0.42**	0.42	0.00	0.41	**0.52**	0.06	0.40	**0.43**	0.00
8	Acc.	0.90	**0.85**	0.80	0.10	**0.85**	0.75	0.65	**0.75**	**0.75**	0.00	**0.90**	0.78	0.80	**0.95**	0.80	0.60	**0.95**	0.73	0.60	**0.90**	0.78	0.50	**0.85**	0.80	0.35
	F_1	0.88	**0.81**	0.79	0.13	**0.83**	0.74	0.71	**0.77**	**0.77**	0.00	**0.84**	0.79	0.75	**0.87**	0.79	0.58	**0.88**	0.74	0.63	**0.88**	0.81	0.56	0.80	**0.84**	0.46
9	Acc.	0.65	0.30	0.35	**0.50**	0.25	0.30	**0.40**	0.20	0.33	**0.80**	**0.40**	0.40	0.00	0.35	**0.38**	0.00	**0.35**	0.28	0.00	**0.40**	0.35	0.10	0.25	0.30	**0.65**
	F_1	0.67	**0.39**	0.31	0.19	**0.31**	0.28	**0.31**	**0.28**	0.34	0.19	**0.47**	0.30	0.00	**0.45**	0.39	0.00	**0.38**	0.24	0.00	**0.44**	0.31	0.12	**0.26**	0.25	0.12
10	Acc.	0.30	0.15	**0.33**	0.00	**0.30**	0.13	0.05	0.00	0.28	0.00	0.30	**0.40**	0.00	**0.30**	0.28	0.00	0.25	**0.33**	0.00	**0.30**	0.30	0.05	**0.25**	0.15	0.05
	F_1	0.24	0.12	**0.27**	0.00	**0.21**	0.07	0.06	0.00	0.22	0.00	0.20	**0.29**	0.00	**0.22**	0.21	0.00	0.18	**0.26**	0.00	**0.31**	0.24	0.06	**0.20**	0.12	0.08
11	Acc.	0.45	**0.40**	0.30	0.00	**0.45**	0.25	0.10	**0.30**	**0.35**	0.00	**0.60**	0.20	0.00	**0.50**	0.33	0.15	**0.55**	0.33	0.00	**0.25**	0.28	0.00	**0.50**	0.23	0.00
	F_1	0.34	**0.31**	0.24	0.00	**0.32**	0.14	0.10	**0.23**	**0.26**	0.00	**0.43**	0.15	0.00	**0.44**	0.25	0.10	**0.39**	0.31	0.00	**0.17**	0.21	0.00	**0.29**	0.19	0.00
12	Acc.	0.45	0.35	**0.40**	0.00	0.25	0.40	**0.35**	**0.55**	0.23	0.00	0.30	0.35	**0.40**	0.20	0.35	**0.90**	0.10	0.38	**0.80**	**0.50**	0.23	0.20	**0.35**	0.20	0.10
	F_1	0.45	0.34	**0.41**	0.00	0.22	**0.45**	0.18	**0.24**	0.24	0.00	**0.25**	0.39	0.09	0.15	0.38	**0.18**	0.10	0.32	0.16	**0.38**	0.25	0.14	**0.23**	0.19	0.02
all	WA	0.54	**0.42**	0.41	0.15	**0.45**	0.33	0.22	0.32	**0.45**	0.10	**0.46**	0.42	0.20	**0.48**	0.47	0.18	**0.46**	0.37	0.16	0.43	**0.44**	0.20	**0.39**	0.37	0.13

Notes: "Bas."/"Rec."/"Occ."/"Ref." denote baseline / reconstructed / training with occluded samples/reference case, Acc, F_1 and WA denote Accuracy, F_1 score and weighted accuracy, respectively. "None" denotes the case without occlusion. LA, RA, LL, RL denote the occlusion of left arm, right arm, left leg, right leg, respectively. Numbers in bold indicate best performance between "Rec."/"Occ."/"Ref.".

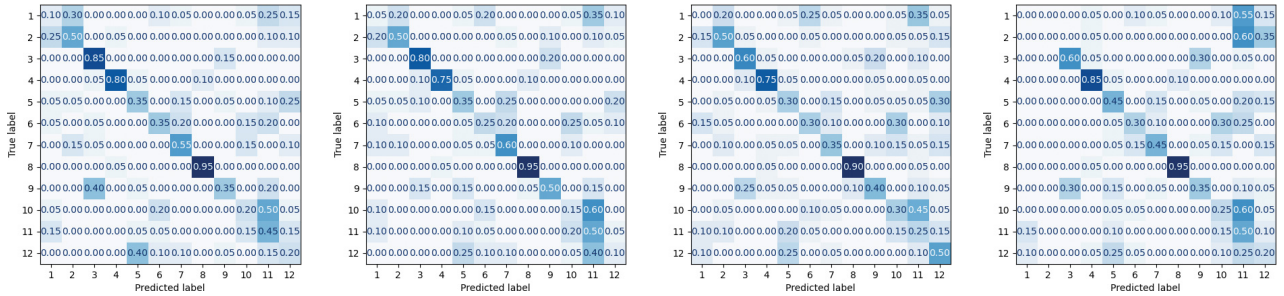
Table 4. Results on UTKinect-Action3D dataset.

			LA			RA			LA+RA			LL			RL			LL+RL			LA+LL			RA+RL		
Class	Metric	None Baseline	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.	Rec.	Occ.	Ref.
0	Acc.	0.50	**0.40**	0.30	0.30	**0.40**	0.30	0.00	**0.40**	0.30	0.00	**0.40**	0.30	0.20	0.40	0.40	**0.50**	**0.40**	0.30	0.20	0.40	0.50	**0.70**	**0.40**	**0.40**	0.00
	F_1	0.48	**0.43**	0.35	0.31	**0.43**	0.35	0.00	**0.39**	0.26	0.00	**0.43**	0.31	0.09	**0.43**	0.41	0.41	**0.43**	0.28	0.07	**0.43**	0.44	0.33	0.39	**0.43**	0.00
1	Acc.	0.70	0.30	0.20	**0.60**	0.10	0.20	**0.30**	0.30	0.20	**0.90**	**0.70**	0.60	0.00	**0.50**	0.30	0.40	**0.60**	0.10	0.00	**0.40**	0.50	0.20	**0.30**	0.30	0.20
	F_1	0.53	0.23	0.16	**0.15**	0.08	0.10	**0.17**	0.13	0.11	**0.18**	**0.57**	0.51	0.00	**0.39**	0.26	0.35	**0.43**	0.10	0.00	**0.32**	0.39	0.06	**0.21**	0.23	0.13
2	Acc.	0.80	**0.80**	0.70	0.00	**0.80**	0.50	0.00	**0.90**	0.80	0.00	**0.80**	0.80	0.10	0.70	**0.80**	0.60	**0.70**	0.60	0.00	**0.90**	**0.90**	0.20	**0.90**	0.40	0.00
	F_1	0.87	**0.76**	0.66	0.00	**0.80**	0.41	0.00	**0.83**	0.76	0.00	**0.80**	0.67	0.13	**0.75**	0.66	0.63	**0.75**	0.46	0.00	0.79	**0.85**	0.27	**0.89**	0.26	0.00
3	Acc.	1.00	**0.30**	0.30	0.00	**0.60**	0.10	0.20	0.00	**0.30**	0.00	**0.70**	0.40	0.40	0.70	0.60	**1.00**	**0.80**	0.30	0.70	0.20	**0.70**	0.00	0.20	**0.30**	**0.30**
	F_1	0.92	0.22	**0.26**	0.00	**0.43**	0.05	0.05	0.00	**0.33**	0.00	**0.58**	0.41	0.40	0.63	0.57	**0.69**	**0.72**	0.18	0.61	0.11	**0.63**	0.00	**0.16**	0.29	0.07
4	Acc.	0.90	0.00	**0.20**	0.00	0.40	0.10	**0.60**	0.00	**0.50**	0.00	0.60	0.20	**0.90**	**0.60**	0.60	0.50	0.60	0.10	**0.80**	0.10	**0.50**	0.00	0.30	0.10	**0.60**
	F_1	0.93	0.00	**0.11**	0.00	**0.40**	0.13	0.14	0.00	**0.41**	0.00	0.57	0.13	**0.73**	**0.60**	0.57	0.39	0.57	0.08	**0.59**	0.08	**0.46**	0.00	**0.20**	0.13	0.14
5	Acc.	1.00	**0.80**	0.50	0.00	**0.80**	0.50	0.00	**0.80**	0.40	0.00	0.80	0.50	**0.90**	**0.80**	0.40	0.70	**0.80**	0.50	**0.80**	**0.80**	0.50	0.10	**0.80**	0.50	0.00
	F_1	0.92	**0.69**	0.49	0.00	**0.69**	0.43	0.00	**0.72**	0.47	0.00	**0.69**	0.50	0.58	**0.69**	0.43	0.65	**0.69**	0.40	0.59	**0.73**	0.60	0.13	**0.76**	0.43	0.00
6	Acc.	1.00	**0.80**	0.50	0.00	**0.80**	0.70	0.00	**0.90**	0.80	0.00	0.90	0.80	**1.00**	0.80	0.80	**0.90**	**0.80**	0.70	0.80	0.80	**1.00**	0.20	**1.00**	0.70	0.00
	F_1	1.00	**0.80**	0.49	0.00	**0.80**	0.66	0.00	**0.93**	0.61	0.00	**0.93**	0.72	0.79	**0.76**	0.68	0.81	**0.76**	0.63	0.51	0.83	**0.86**	0.13	**0.96**	0.63	0.00
7	Acc.	0.40	0.50	**0.60**	**0.70**	**0.40**	0.30	0.00	0.20	**0.60**	0.00	**0.40**	0.30	0.00	**0.20**	0.70	0.10	0.30	**0.40**	0.00	**0.40**	0.30	**0.40**	0.40	**0.50**	0.00
	F_1	0.43	**0.25**	0.22	0.23	**0.28**	0.21	0.00	0.17	**0.35**	0.00	**0.40**	0.19	0.00	**0.12**	0.54	0.10	0.27	**0.34**	0.00	**0.24**	0.21	0.11	**0.34**	0.28	0.00
8	Acc.	0.60	0.40	**0.50**	0.00	**0.40**	0.30	0.00	0.20	**0.50**	0.00	0.40	0.60	**0.50**	**0.40**	0.60	**0.40**	0.40	**0.50**	0.30	0.30	**0.50**	0.20	**0.40**	**0.40**	0.00
	F_1	0.52	0.36	**0.44**	0.00	**0.36**	0.29	0.00	0.16	**0.47**	0.00	0.27	0.55	**0.40**	0.36	0.55	**0.43**	0.29	**0.53**	0.27	0.26	**0.53**	0.27	**0.36**	0.44	0.00
9	Acc.	1.00	**0.70**	0.40	0.00	**1.00**	0.30	0.40	0.30	**1.00**	0.00	**1.00**	0.70	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.90	**1.00**	0.80	**1.00**	0.20	**1.00**	0.80	0.00
	F_1	0.87	**0.52**	0.23	0.00	**0.83**	0.10	0.47	0.13	**0.88**	0.00	**0.87**	0.59	0.69	0.87	**0.96**	**0.96**	**0.87**	0.81	0.74	0.58	**0.92**	0.07	**0.87**	0.76	0.00
all	WA	0.79	**0.50**	0.42	0.16	**0.57**	0.33	0.15	0.40	**0.54**	0.09	**0.67**	0.52	0.50	**0.**61	**0.62**	0.61	**0.64**	0.44	0.46	0.51	**0.64**	0.22	**0.57**	0.44	0.11

Notes: "Bas."/"Rec."/"Occ."/"Ref." denote baseline/reconstructed/training with occluded samples/reference case, Acc, F_1 and WA denote Accuracy, F_1 score and weighted accuracy, respectively. "None" denotes the case without occlusion. LA, RA, LL, RL denote the occlusion of left arm, right arm, left leg, right leg, respectively. Numbers in bold indicate best performance between "Rec."/"Occ."/"Ref.".

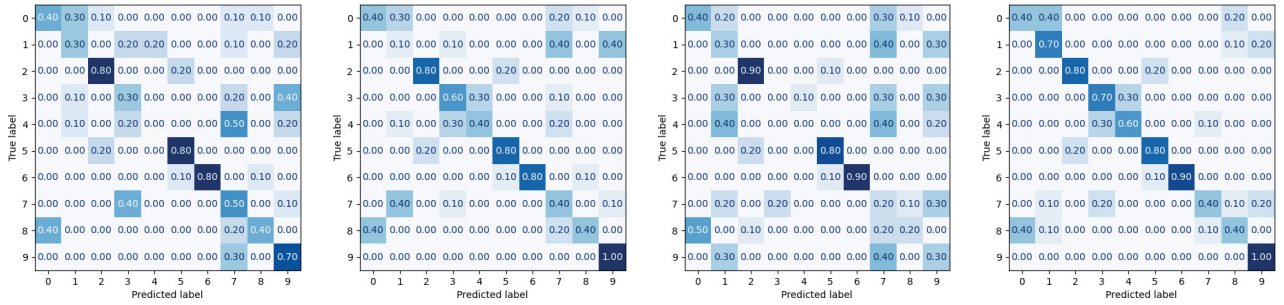


(a) Left Arm (b) Right Arm (c) Left Arm & Right Arm (d) Left Leg

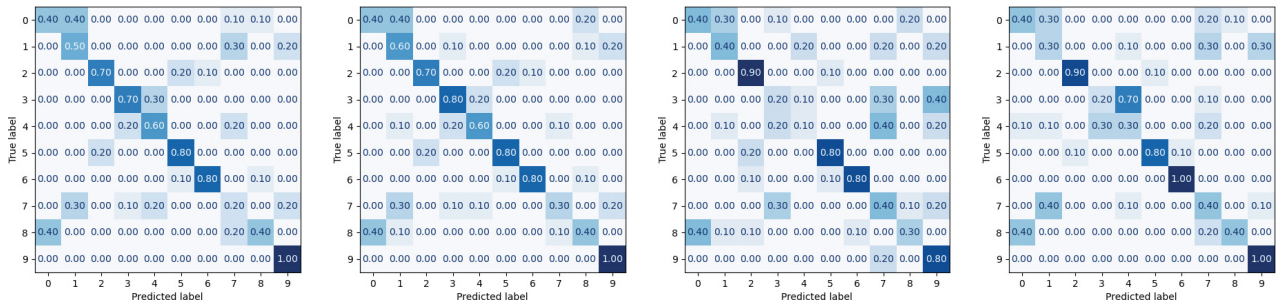


(e) Right Leg (f) Left Leg & Right Leg (g) Left Arm & Left Leg (h) Right Arm & Right Leg

Fig. 9. Average confusion matrices from five trials, for classification for the SYSU-3D-UOI dataset upon removing the body part(s) denoted in the caption of the corresponding subfigure.



(a) Left Arm (b) Right Arm (c) Left Arm & Right Arm (d) Left Leg



(e) Right Leg (f) Left Leg & Right Leg (g) Left Arm & Left Leg (h) Right Arm & Right Leg

Fig. 10. Average confusion matrices from five trials, for classification for the UTKinect-Action-3D dataset upon removing the body part(s) denoted in the caption of the corresponding subfigure.

In case of the UTKinect-Action-3D dataset, the WA was 0.79 without and body part removal, and ranged between 0.09–0.61 without any body part removal and ranged between 0.40–0.67 upon reconstruction. Moreover, reconstruction outperformed occlusion in 5 out of 8 cases. Also, in that case, in 7 out of 8 cases, significant improvement was observed, in terms of WA, while performance was equal in case of removal of right arm. Since activities used to evaluate the proposed approach mainly consisted of upper body motion, also in the case of the UTKinect-Action-3D dataset, in the results of Table 4 it could be observed that in all the remaining cases of occluded arms, the improvement of WA is large, with the most notable examples the cases of both arms, wherein WA improves from 0.09 to 0.40 and right arm/right leg wherein WA improves from 0.11 to 0.57. Upon careful observation of the confusion matrices depicted in Fig. 10 and considering the classification results without occlusion of any part, for each occlusion case we should notice that occlusion mainly affects activities *pickUp*, *carry*, *pull* and *clapHands*, although most activities that are based on the motion of arms are affected.

In all cases, occlusion of some body parts leads to a severe drop of performance; in many occasions, accuracy and F_1 score of some classes dropped to zero/near zero values. In some of these cases, reconstruction was successful, leading to near perfect recognition. Interestingly, addition of occluded samples within the training set, in certain cases leads to an improved performance compared to the baseline case. This has been observed in all four datasets and we believe that this is an effect of the augmentation of the training data set that we have previously mentioned. Finally, in several activities and for specific cases of occlusion, it is observed that classification of the reconstructed samples exhibits slightly inferior performance compared to the occluded samples. It is our belief that in those cases, the specific body part which is occluded is less "relevant" for those activities, although in some cases the reconstruction approach may fail and lead to misleading joint positions, as in the example of Fig. 2(b).

5. Conclusions and Future Work

In this paper, we presented an approach for human activity recognition under partial occlusion, i.e. occlusion of one/two parts. Our approach was based on the motion of 3D skeletal joints and occlusion was simulated upon manually removing structured sets of joints which corresponded to arms and/or legs. We managed to reconstruct the missing joints by formulating the problem as a regression task. Specifically, we trained a convolutional recurrent neural network using occluded and nonoccluded samples and then used it to reconstruct missing joints of occluded samples. For classification of reconstructed samples, we used a long short term memory network. We then evaluated the proposed approach using four publicly available datasets. Among them, we experimented with several different types of activities, such as simple activities, activities of daily living, medical-related activities and interactions with objects. We showed that this way, we could achieve a significant boost of performance in a classification task, when using reconstructed samples instead of occluded ones, almost in any activity.

The novelties of our work were as follows: (a) we formulated the problem of missing skeleton joints due to occlusion as a regression problem, within a human motion activity recognition task and implemented a novel deep neural network architecture to reconstruct missing skeletal joints; (b) contrary to other state-of-the-art works, we did not use occluded samples within the training process; (c) the occluded parts were not visible from any camera, even in case of a multi-camera setup.

The proposed approach could benefit several HAR tasks which take place into dynamic environments and are prone to occlusion. Apart from applications within ambient assisted environment i.e. where the goal is to detect activities of daily living (ADLs),[27] or medical-related events that require some kind of attention or intervention (e.g. staggering, falling, etc.). In such scenarios, as we have already mentioned, occlusion may take place due to e.g. furniture or other people present. Moreover, it could also be of great utilization in AR environments and applications, since human motion is tied to the overall user experience and, HAR may act as a means of assessing user engagement e.g. when a visitor of a museum makes a phone call while interacting with an AR application, this should be an indicator of low engagement, while when she/he is reading in front of an AR screen, this should be an indicator of high engagement. Moreover, another important field of application is the one of surveillance,

where the goal is to identify certain types of unwanted activity or abnormal events. In cases where a subject acts in public places, occlusion may take place due to obstacles or the presence of other people/crowd.

Future research work may target several aspects of the problem of HAR, focusing on occlusion. Our regression approach could be enhanced and improved e.g. by replacing the interpolation step herein used with a temporal augmentation approach such as the one that has been proposed by Kwon *et al.*[33]. Moreover, we would like to experiment with other state-of-the-art architectures such as transformers[34] or techniques that work with sequences, such as seq2seq[35] or time-series such as multivariate CNNs.[36] Also, the use of modern classifiers could be investigated for classification. For example, other possible research directions may include experiments with approaches such as Neural Dynamic Classification,[37] Dynamic Ensemble Learning,[38] Finite Element Machine for fast learning[39] or self-supervised learning.[40] Additionally, in the training phase we will further investigate the inclusion of occluded samples upon reconstruction.

Also, when dealing with the occlusion aspect of HAR, we would like to investigate cases such as temporally partial occlusion. Then we would like to investigate the use of other deep neural network architectures, such as generative adversarial networks (GANs), e.g. for generating the missing body parts. Also, since in previous work we have showed that handcrafted features are able to boost recognition performance of deep approaches,[41] it would be interesting to experiment with other methodologies for feature extraction that are based on the geometry and the motion of skeletons, as the one proposed by Avola *et al.*[42] Moreover, we would like to perform experiments using larger datasets. We would like to perform comparisons of the given approach to one that uses occluded samples for training the neural network that we have herein used for classification, without a regression step.

Finally, we plan to perform real-life experiments into an assistive living environment or simulate occlusions occurring within such an environment by constructing masks using 3d models of real objects.[43] In that case and for privacy preservation issues, pose estimation approaches that do not depend on visual could replace camera-based approaches. A possible candidate could be based on the use of Wi-Fi signals,[44] since Wi-Fi routers are typically encountered within any home environment.

Acknowledgments

The implementation of the doctoral thesis was co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme Human Resources Development, Education and Lifelong Learning in the context of the Act "Enhancing Human Resources Research Potential by undertaking a Doctoral Research" Sub-action 2: IKY Scholarship Programme for Ph.D. candidates in the Greek Universities. This research was funded by the European Union and Greece (Partnership Agreement for the Development Framework 2014–2020) under the Regional Operational Programme Ionian Islands 2014–2020, project title: "Indirect costs for the project 'Smart digital applications and tools for the effective promotion and enhancement of the Ionian Islands biodiversity, project number: 5034557'".

References

1. E. Mathe, A. Maniatis, E. Spyrou and P. Mylonas, A deep learning approach for human action recognition using skeletal information. In *GeNeDis 2018. Advances in Experimental Medicine and Biology*, P. Vlamos (eds), Vol. 1194 (Springer, Cham, 2020), https://doi.org/10.1007/978-3-030-32622-7_9.
2. W. Jiang and Z. Yin, Human activity recognition using wearable sensors by deep convolutional neural networks, *23rd ACM Int. Conf. Multimedia* (Brisbane, Australia, 2015), pp. 1307–1310.
3. S. Lee, D.-W. Lee and M. S. Kim, A deep learning-based semantic segmentation model using MCNN and attention layer for human activity recognition, *Sensors* **23**(4) (2023) 2278.
4. S. Zhu, W. Chen, F. Liu, X. Zhang and X. Han, Human activity recognition based on a modified capsule network, *Mob. Inf. Syst.* **2023** (2023) 8273546.
5. J. Cheng, M. Sundholm, B. Zhou, M. Hirsch and P. Lukowicz, Smart-surface: Large scale textile pressure sensors arrays for activity recognition, *Pervasive Mob. Comput.* **30** (2016) 97–112.
6. S. Majumder, T. Mondal and M. J. Deen, Wearable sensors for remote health monitoring, *Sensors* **17**(1) (2017) 130.

7. A. Keogh, J. F. Dorn, L. Walsh, F. Calvo and B. Caulfield, Comparing the usability and acceptability of wearable sensors among older Irish adults in a real-world context: Observational study, *JMIR mHealth uHealth* **8**(4) (2020) e15704.
8. P. Wang, W. Li, P. Ogunbona, J. Wan and S. Escalera, RGB-D-based human motion recognition with deep learning: A survey, *Comput. Vis. Image Underst.* **171** (2018) 118–139.
9. S. Ranasinghe, F. Al Machot and H. C. Mayr, A review on applications of activity recognition systems with regard to performance and evaluation, *Int. J. Distrib. Sens. Netw.* **12**(8) (2016) doi: 10.1177/1550147716665520.
10. S. Antoshchuk, M. Kovalenko and J. Sieck, Gesture recognition-based human–computer interaction interface for multimedia applications, *Digitisation of Culture: Namibian and International Perspectives* (Springer, 2018), pp. 269–286.
11. A. Papadakis, E. Mathe, E. Spyrou and P. Mylonas, A geometric approach for cross-view human action recognition using deep learning, *IEEE Int. Symp. Image and Signal Processing and Analysis (ISPA)* (Dubrovnik, Croatia, 2019), pp. 258–263.
12. Z. Zhang, Microsoft kinect sensor and its effect, *IEEE Multimed.* **19**(2) (2012) 4–10.
13. I. Giannakos, E. Mathe, E. Spyrou and Ph. Mylonas, A study on the effect of occlusion in human activity recognition, *14th PErvasive Technologies Related to Assistive Environments Conf.* (Corfu, Greece, 2021), pp. 473–482.
14. Y. Du, Y. Fu and L. Wang, Skeleton based action recognition with convolutional neural network, *IAPR Asian Conf. Pattern Recognition (ACPR)* (IEEE, 2015), pp. 579–583.
15. Y. Hou, Z. Li, P. Wang and W. Li, Skeleton optical spectra-based action recognition using convolutional neural networks, *IEEE Trans. Circuits Syst. Video Technol.* **28**(3) (2016) 807–811.
16. Q. Ke, S. An, M. Bennamoun, F. Sohel and F. Boussaid, Skeletonnet: Mining deep part features for 3-d action recognition, *IEEE Signal Process. Lett.* **24** (6) (2017) 731–735.
17. D. Koutrintzes, E. Spyrou, E. Mathe and Ph. Mylonas, A multimodal fusion approach for human activity recognition, *Int. J. Neural Syst.* **33**(1) (2023) 2350002.
18. C. Li, Y. Hou, P. Wang and W. Li, Joint distance maps based action recognition with convolutional neural networks, *IEEE Signal Process. Lett.* **24**(5) (2017) 624–628.
19. M. Liu, H. Liu and C. Chen, Enhanced skeleton visualization for view invariant human action recognition, *Pattern Recognit.* **68** (2017) 346–362.
20. P. Wang, Z. Li, Y. Hou and W. Li, Action recognition based on joint trajectory maps using convolutional neural networks, in *Proc. 2016 ACM on Multimedia Conf.* (Amsterdam, the Netherlands, 2016).
21. A. Iosifidis, A. Tefas and I. Pitas, Multi-view human action recognition under occlusion based on fuzzy distances and neural networks, *European Signal Processing Conf. (EUSIPCO)* (IEEE, 2012), pp. 1129–1133.
22. R. Gu, G. Wang and J. N. Hwang, Exploring severe occlusion: Multi-person 3d pose estimation with gated convolution, *Int. Conf. Pattern Recognition (ICPR)* (IEEE, 2021), pp. 8243–8250.
23. T. Liu, J. J. Sun, L. Zhao, J. Zhao, L. Yuan, Y. Wang, L.-C. Chen, F. Schroff and H. Adam, View-invariant, occlusion-robust probabilistic embedding for human pose, *Int. J. Comput. Vis.* **130**(1) (2020) 111–135.
24. F. Angelini, Z. Fu, Y. Long, L. Shao and S. M. Naqvi, 2d pose-based real-time human action recognition with occlusion-handling, *IEEE Trans. Multimed.* **22**(6) (2019) 1433–1446.
25. I. A. Kostis, E. Mathe, E. Spyrou and Ph. Mylonas, Human activity recognition under partial occlusion, *23rd Int. Conf. Engineering Applications of Neural Networks:, EAAAI/EANN* (Chersonisos, Crete, Greece, 2022), pp. 297–309.
26. C. Liu, Y. Hu, Y. Li, S. Song and J. Liu, PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding, arXiv:1703.07475.
27. M. P. Lawton and E. M. Brody, Assessment of older people: Self-maintaining and instrumental activities of daily living, *The Gerontologist* **9**(3 Part 1) (1969) 179–186.
28. A. Shahroudy, J. Liu, T. T. Ng and G. Wang, NTU RGB+ D: A large scale dataset for 3d human activity analysis, *IEEE Conf. Computer Vision and Pattern Recognition* (Las Vegas, Nevada, USA, 2016).
29. J. F. Hu, W. S. Zheng, J. Lai and J. Zhang, Jointly learning heterogeneous features for RGB-D activity recognition, *IEEE Conf. Computer Vision and Pattern Recognition* (Boston, MA, USA, 2015), pp. 5344–5352.
30. L. Xia, C. C. Chen and J. K. Aggarwal, View invariant human action recognition using histograms of 3d joints, *2012 IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops* (Providence, RI, USA, 2012), pp. 20–27.
31. F. Chollet and others, Keras. https://github.com/fchollet/keras (2015).
32. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudeva, P. Warden, M. Wicke, Y. Yu and X. Zheng, TensorFlow: A system for Large-Scale machine learning, *12th USENIX Symp. Operating Systems Design and Implementation (OSDI 16)* (Savannah, GA, USA, 2016), pp. 265–283.
33. T. Kwon, B. Tekin, S. Tang and M. Pollefeys, Context-aware sequence alignment using 4D skeletal augmentation, *IEEE/CVF Conf. Computer Vision*

and Pattern Recognition (New Orleans, Louisiana, USA, 2022), pp. 8172–8182.
34. B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen and M. Wu, Two-stream convolution augmented transformer for human activity recognition, *AAAI Conf. Artificial Intelligence*, Vol. 35(1) (Virtual, 2021), pp. 286–293.
35. I. Sutskever, O. Vinyals and Q. V. Le, Sequence to sequence learning with neural networks, *Adv. Neural Inf. Process. Syst.* **27** (2014) 1–9.
36. C. L. Liu, W. H. Hsaio and Y. C. Tu, Time series classification with multivariate convolutional neural network, *IEEE Trans. Ind. Electron.* **66**(6) (2018) 4788–4797.
37. M. H. Rafiei and H. Adeli, A new neural dynamic classification algorithm, *IEEE Trans. Neural Netw. Learn. Syst.* **28**(12) (2017) 3074–3083.
38. K. M. Alam, N. Siddique and H. Adeli, A dynamic ensemble learning algorithm for neural networks, *Neural Comput. Appl.* **32**(12) (2020) 8675–8690.
39. D. R. Pereira, M. A. Piteri, A. N. Souza, J. P. Papa and H. Adeli, FEMa: A finite element machine for fast learning, *Neural Comput. Appl.* **32**(10) (2020) 6393–6404.
40. M. H. Rafiei, L. V. Gauthier, H. Adeli and D. Takabi, Self-supervised learning for electroencephalography, *IEEE Trans. Neural Netw. Learn. Syst.* (2022).
41. D. Koutrintzes, E. Mathe and E. Spyrou, Boosting the performance of deep approaches through fusion with handcrafted features, *Int. Conf. Pattern Recognition Applications and Methods — ICPRAM* (Virtual, 2022), pp. 370–377.
42. D. Avola, M. Cascio, L. Cinque, G. L. Foresti, C. Massaroni and E. Rodol, 2D Skeleton-based action recognition via two-branch stacked LSTM-RNNs, *IEEE Trans. Multimed.* **22**(10) (2020) 2481–2496.
43. K. Peng, A. Roitberg, K. Yang, J. Zhang and R. Stiefelhagen, Delving deep into one-shot skeleton-based action recognition with diverse occlusions, *IEEE Trans. Multimed.* **25** (2023) 1489–1504.
44. D. Avola, M. Cascio, L. Cinque, A. Fagioli and G. L. Foresti, Human silhouette and skeleton video synthesis through Wi-Fi signals, *Int. J. Neural Syst.* **32**(5) (2022) 2250015.
45. I. Vernikos, E. Mathe, A. Papadakis, E. Spyrou and P. Mylonas, An image representation of skeletal data for action recognition using convolutional neural networks, *ACM Int. Conf. PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece, 2019), pp. 325–326.