A study on the Effect of Occlusion in Human Activity Recognition

ILIAS GIANNAKOS, Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", Greece

EIRINI MATHE, Department of Informatics, Ionian University, Greece and Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", Greece

EVAGGELOS SPYROU, Department of Informatics and Telecommunications, University of Thessaly, Greece and Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", Greece

PHIVOS MYLONAS, Department of Informatics, Ionian University, Greece

The problem of occlusion plays a crucial role in real-life human activity recognition applications. However, most research works either underestimate it, or base their training solely on datasets shot under laboratory conditions, i.e., without any partly or full occlusion. In this work we perform a study on the effect of occlusion in the task of human activity recognition and the domains of the recognition of a) activities of daily living; and b) medical conditions. Throughout our experiments we use a convolutional neural network that has been trained using a 2D representation of skeleton motion for all available joints, i.e., without using any occluded samples. We evaluate our approach using two challenging, publicly available human motion datasets upon removing one or more body parts.

CCS Concepts: • **Computing methodologies** → **Activity recognition and understanding**; **Neural networks**.

Additional Key Words and Phrases: convolutional neural networks, recognition of activities of daily living, recognition of medical conditions, assistive living

ACM Reference Format:
Ilias Giannakos, Eirini Mathe, Evaggelos Spyrou, and Phivos Mylonas. 2021. A study on the Effect of Occlusion in Human Activity Recognition . 1, 1 (March 2021), 14 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Advances in the fields of medicine have allowed people to live longer. More specifically, during the last few years, life expectancy in Europe has increased from 62 years in 1950 to 77.8 years in 2015 [24]. Globally, there are approx. 727M individuals aged 65 years or more; this number is expected to double until 2050 [26]. Accordingly, the corresponding share of population is expected to increase from 9.3% to approx. 16%. Moreover, the vast majority of older adults prefer their own home, over staying within nursing facilities. In USA, the 77% of age 40 and older population would prefer to receive care in their home, while only 4% in a nursing home [20]. Also, the costs of nursing homes are increasing,

Authors' addresses: Ilias Giannakos, Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", Athens, Greece, igiannakos@iit.demokritos.gr; Eirini Mathe, Department of Informatics, Ionian University, Corfu, Greece, Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", Athens, Greece, emathe@iit.demokritos.gr; Evaggelos Spyrou, Department of Informatics and Telecommunications, University of Thessaly, Lamia, Greece, Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", Athens, Greece, espyrou@uth.gr; Phivos Mylonas, Department of Informatics, Ionian University, Corfu, Greece, fmylonas@ionio.gr.

therefore in many cases it may not be an affordable option. Therefore, there exists a continuously increasing need for the development of assistive technologies for in-home living.

Recognition of activities of daily living (ADLs) [14] consist one of the main tasks within the broader research area of assistive living [23]. Those activities within a healthcare context, commonly refer to daily self-care activities. The ability/inability of an individual to perform ADLs may be used as some means of measurement of their functional status. Common ("basic") ADLs include feeding without help, personal hygiene, homemaking, dressing etc. Moreover "instrumental" ADLs are not necessary for fundamental functions, yet they are necessary so that the individual could be able to be independent. Such ADLs include e.g., cleaning, cooking, using the telephone etc. Therefore, within an assistive living scenario, in order to assess the ability of a person to live independently, it is necessary to verify if/when a given set of ADLs takes place.

Of course, this recognition requires a set of sensors installed either on the subject's environment or worn by the user, thus it consists one of the hottest topics in the area of pervasive computing. The former typically include video/thermal cameras, microphones, infrared, pressure, magnetic, RFID sensors [5] etc. Their role is to capture motion, speech, sound events, presence within some space, interaction with objects etc. On the other hand the latter include smartwatches, smartphones, RFIDs, hand worn and vital monitoring sensors, to capture monitor, presence within a space, vital signs, gestures etc. [5]. All available measurements are collected and processed so as to recognize a predefined set of ADLs and draw conclusions regarding the subject's state. In many scenarios, the recognition of ADLs is combined with the recognition of several simple "medical" events such as e.g., coughing, chest pain, staggering, falling etc.

Since wearable sensors is [21] are non-invasive, while offering an efficient, low-cost solution, it has been shown that in many cases elder subjects do not intent to wear them, apart from occasions when it is necessary; also their usability is below average [12]. Therefore, in many approaches that aim to provide low-cost solutions without the use of wearables and without overloading the subject's space with a plethora of sensors, the use of only cameras that capture the subjects' motion and aim to accordingly detect appropriate ADLs is preferred. However, it is well-known that camera-based approaches suffer from viewpoint and illumination changes and also from occlusion.

In previous work [22] we have proposed an approach for human activity recognition (HAR) focusing on ADLs and using only visual data. Our approach was based on 3D skeletal motion of human joints which was extracted upon processing of RGB and depth data modalities. We evaluated our approach under different viewpoints and showed that the recognition of ADLs is still feasible, yet accuracy decreases, as expected. In this work our goal is to assess how partial occlusion of the subject affects the accuracy of recognition. We simulate occlusion by removing parts of captured visual data and we evaluate using visual data comprising of ADLs and medical conditions from two publicly available challenging datasets. Note that we use models that have not been trained with activity instances that have been affected by occlusion. To the best of our knowledge, our work is the first to perform such an evaluation.

The rest of this work is organized as follows: section 2 presents research works that aim to assess or even tackle the effect of occlusion in HAR-related scenarios. Then, in section 3 we present the proposed methodology for recognition and the approach we followed for simulating occlusion. Results of our approach are presented in section 4. Finally, conclusions are drawn in section 5, wherein plans for future work are also presented.

2 RELATED WORK

The problem of HAR from video sequences may be divided into two major tasks, as stated by Wang et al. [28]: a) segmented; and b) continuous recognition. Within the former category it is assumed that the video at hand contains exactly one action to be recognized, i.e., any poses/motion before/after the activity have been previously removed

upon a segmentation/trimming process. Within the latter, the goal is to recognize activities within a given video, that may contain an unlimited number of activities, or even none. Therein, temporal localization of actions is typically a necessary pre-processing step.

Although several approaches had been proposed during the early years of HAR, they were typically limited to small numbers of simple actions, while being prone to drop of performance due to e.g., viewpoint changes and/or illumination changes and/or occlusion. During the last decade, the rise of deep neural network architectures has been the main cause for growing research within the broader area of HAR. Therein, the main open challenges may include the representation, the analysis and the recognition of the actions [3], while a plethora of applications have benefited.

The two main deep architectures that have been used in the field of HAR are Convolutional Neural Networks (CNNs) [17] and Recurrent Neural Networks (RNNs) [7]. Note that when CNNs are used, as within our approach, typically a 2D image representation of RGB/depth or skeletal sequences is required so as to be used as input. Obviously, when building such a representation, the goal is to capture spatio-temporal information of motion and reflect it to the color and/or texture properties of the representation. Of course, in such approaches an intermediate hand-crafted feature extraction step is typically omitted within the process. Our approach is based on skeletal data, which typically consist of a set of skeletal joints moving in 3D space over time, i.e., for each joint 3 1-D signals are generated per action. Typically, the extraction of joints from video requires RGB and depth information. To capture both modalities, one popular off-the-shelf solution is the Microsoft Kinect v2 sensor, which combines an RGB and a Time-of-Flight camera and provides a powerful software development kit (SDK) for the extraction of joint motion. We should note that skeletons are prone to errors; the most important causes are occlusion and viewpoint changes.

During the last few years, a plethora of research works for HAR based on 2D representations of skeletal data have been presented. Du et al. [6] grouped the set of extracted joints into five groups, i.e., arms, legs and the trunk and created pseudo-colored image sequences to capture spatio-temporal information; each color component was formed by one of the spatial coordinates. In another representation proposed by Wang et al. [29], "joint trajectory maps" were created based on joint trajectories and appropriately setting saturation and brightness. Similarly, "skeleton optical spectra" were proposed by Hou et al. [9], wherein hue was set based on temporal variation of joint motion. "Joint distance maps" have been proposed by Li et al. [18], encoding pair-wise joint distances, wherein hue was used to encode distance variations. Other approaches such as the one of Liu et al. [19] enhanced joint representations with extra information, i.e., time and joint label. Few approaches such as the one of Ke et al. [11] first extract hand-crafted features and then use them to generate image representations.

Although occlusion consists one of the major causes of performance drop in HAR, few are the research efforts that have dealt with assessing its effect or overcoming it. In the work of Iosifidis et al. [10], a multi-camera setup is used for recognition. For simulation of occlusion, they first trained their algorithm using all available cameras and then tested using a randomly chosen subset. However, in all cases the remaining cameras are able to capture the whole body of the subjects. In the work of Gu et al. [8], randomly generated occlusion masks are used in both training and testing. Each mask covers more than one 2D joints. Liu et al. [15] study two augmentation strategies for modelling the effect of occlusion. The first discards independent keypoints, while the second discards structured keypoints that compose main body parts. Note that occluded samples are included in the training process. Similarly, Angelini et al. [2] 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.

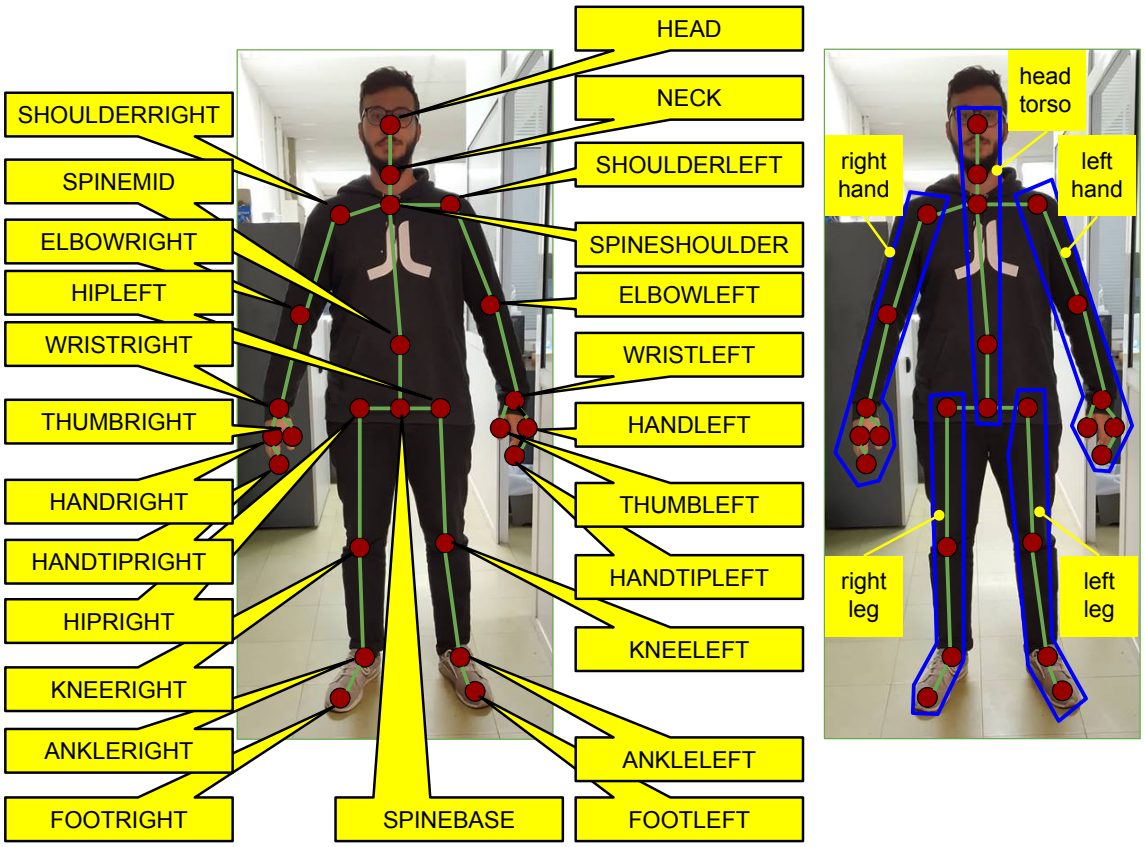


Fig. 1. Left: The 25 skeletal joints extracted by Microsoft Kinect; Right: the joints divided into five main body parts.

3 METHODOLOGY

3.1 Extraction and Pre-processing of Skeletal Data

For recognition of human activities, our approach is based on the extraction of trajectories of skeletal joints, as they move within the 3D space, when an action is performed, over time. More specifically, we require the position of each joint co-ordinate, i.e., $x(t)$, $y(t)$ and $z(t)$. Such skeletal data are typically calculated using RGB and depth video sequences. A popular approach, which we adopt in this work is to use skeletal sequences extracted by the Microsoft Kinect SDK [13]. More specifically, we use data that have been captured using the Microsoft Kinect v2 sensor. These data consist of 25 human joints for each skeleton, which are organized as a 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. An example of a skeleton is illustrated in Fig. 1. Note that for reasons that will clarify in subsection 3.3, in this figure joints are also shown as being grouped to form meaningful body parts, i.e., arms, legs and the torso.

We consider each joint co-ordinate as an 1-D signal. Therefore, upon using all 25 joints, with 3 coordinates each, 75 such signals result for any given video sequence. Upon observing activities as performed by real human subjects, we may observe the following: a) duration of activities varies as different activities may require different amounts of time; b)

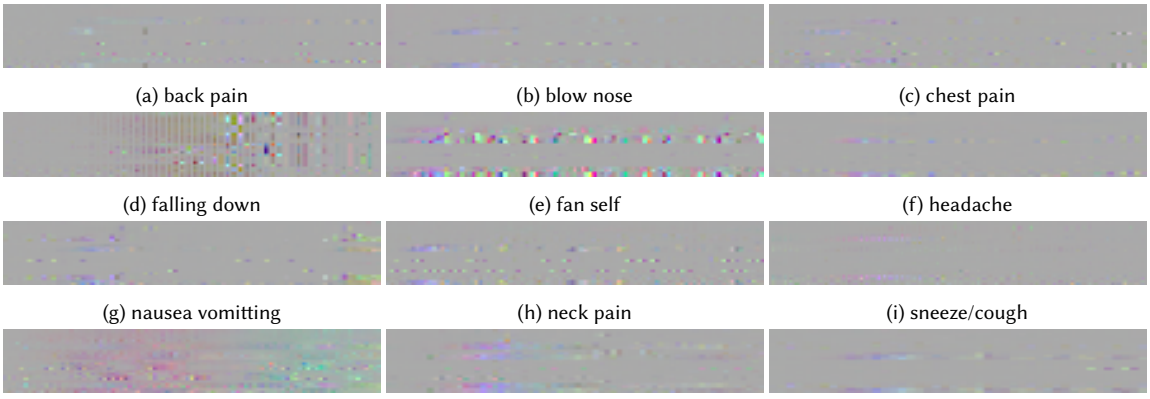


Fig. 2. Sample pseudo-colored images created for the 12 medical classes of NTU RGB+D dataset. Figure best viewed in color.

when different subjects perform the same activity, their duration also varies; c) the same subject may perform the same activity with varying duration. Therefore, to address this temporal variability we impose a linear interpolation step, setting the number of frames within all videos equal to $N = 150$. This frame number has been heuristically defined, upon a series of initial experiments. Moreover, we assume that our approach falls to the category of *segmented* recognition, i.e., we consider the problem of activity localization within a video as "solved," i.e., we work using pre-segmented video sequences that contain exactly one activity to be recognized.

3.2 Recognition of Activities

Upon the aforementioned pre-processing of skeletal data and the creation of joint motion sequences, the next step is to create a visual representation, which could be used for training of a CNN. Similarly to approaches that have been presented in brief in section 2 and continuing our previous work that has been introduced in [27], we opted for pseudo-colored images that aim to capture inter-joint distances during an action, using the 3D joint trajectories. More specifically our approach works as follows:

From each video sequence, we calculate coordinate differences between consecutive frames. To create the pseudo-colored images, x, y, z coordinates correspond to Red (R), Green (G) and Blue (B) color channels, respectively. By $x_i(n)$ we denote the x-position of the i-th joint and within the n-th frame. For example, let us consider R denote the R channel of the pseudo-colored image. The value of a given pixel $R(i, n)$ is calculated as the difference of the positions of this joint in the $x-$axis for frames n, $n + 1$, i.e., for consecutive frames. Therefore $R(i, n) = x_i(n + 1) - x_i(n)$, where $i = 1, \ldots, N$. Similarly, B and G channels are constructed. We argue that the way these pseudo-colored images are formed, both the temporal and the spatial properties of the skeleton trajectories are preserved. Examples of the created images are presented in Figs. 2 and 3.

The CNN architecture we used for classification in short is as follows: The first 2D convolutional layer filters the 25×150 input pseudo-colored image with 5 kernels of size 3×3. The first pooling layer uses max-pooling to perform 2×2 subsampling. Then the second convolutional 2D layer filters the resulting image with 10 kernels of size 3×3 followed by a second polling layer that also uses max-pooling to perform 2×2 subsampling. A third and a fourth convolutional layer follow with 10 and 15 filter kernels, respectively; the size of each being 3×3. Then, the last max-pooling layer follows,

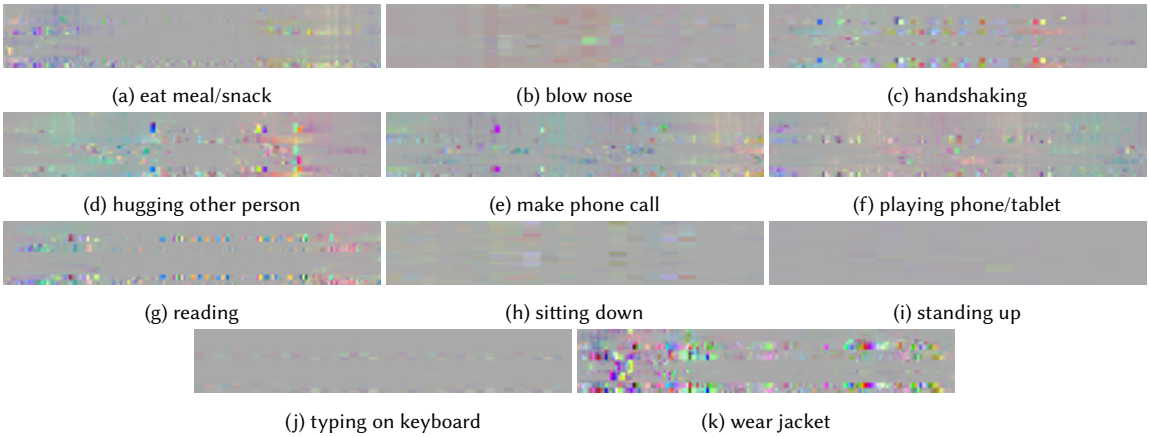


Fig. 3. Sample pseudo-colored images created for 11 classes that are related to ADLs of PKU-MMD dataset. Figure best viewed in color.

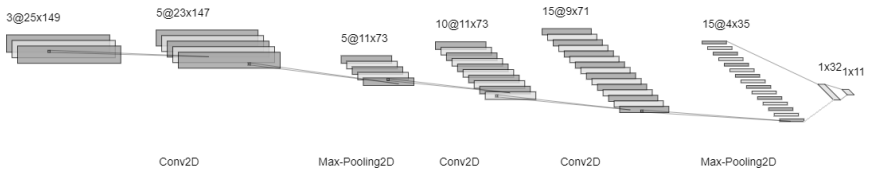


Fig. 4. The convolutional neural network that has been used throughout our experiments.

performing another 2×2 subsampling. Afterwards, a flatten layer transforms the output of the last pooling layer into a vector, which consists the input to a dense layer upon applying a dropout layer with dropout rate 0.5. Finally, a second dense layer produces the output of the network. The architecture of the CNN is illustrated in Fig. 4.

3.3 Simulation of Occlusion

In real-life scenarios, occlusion is probably the most important factor that compromises the performance of many HAR approaches. In the context of assistive living, occlusion may occur mainly due to e.g., furniture or the presence of more than one people in the same room. As expected, it results to loss of visual information regarding the subject's posture and motion, which in many cases may be crucial for the accurate recognition of several activities. Of course, many activities consist of the motion of one/two or even more body parts, thus partial occlusion may prevent their recognition.

As we have already mentioned, in this work our primary goal is to assess the effect of occlusion within a HAR approach. Most public datasets such as the PKU-MMD dataset [16], and the NTU RGB+D [25], which we herein use for evaluation, have been recorded in laboratory conditions. This means that illumination is controlled, while occlusion is prevented. Therefore, to create occluded activity samples and similarly to [8] we discard structured skeleton joints, i.e., subsets that correspond to a body part. Moreover, we assume that occlusion is not temporary, i.e., one or more parts remain occluded during the whole duration of the activity. Contrary to [10], the whole skeleton is never "visible." Also, contrary to [2; 8; 15] we by no means include any occluded samples within the training process.

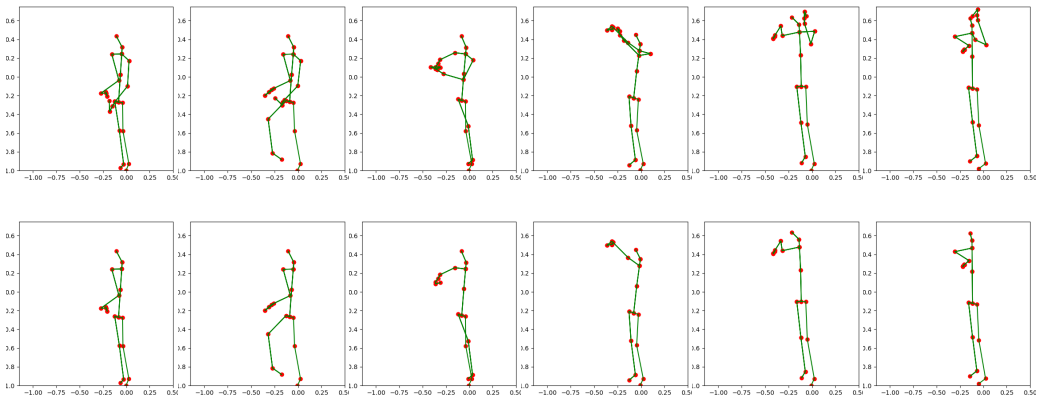


Fig. 5. Example skeleton sequences of the activity *stretch oneself* from the NTU-RGB+D dataset. First row: skeletons include all 25 joints; Second row: joints corresponding to left arm have been discarded.

As we have already mentioned, we use skeletons extracted by the Kinect v2 camera, which consist of 25 joints. We group these joints to form 5 body parts, as illustrated in Fig. 1. More specifically, we consider hands, legs and the torso. Each hand comprises of shoulder, elbow, wrist, hand and hand-tip. Moreover, each leg comprises of hip, knee, ankle and foot. Finally, the torso comprises of head, neck, spine-shoulder, spine-mid and spine-base.

We use the same trained architecture, described in section 3.2. For testing, we consider the cases where one/two arms are occluded, one/two legs are occluded and an arm and a leg. Intuitively, when two parts are occluded, the most expected case is that they both are from the same side. Moreover, our initial experiments indicated that by removing the torso the accuracy was not significantly affected. Therefore, throughout our experiments, torso is always present, while one arm, one leg, both arms, both legs or an arm and a leg from the same side may be missing. An example of an activity with and without occlusion of a body part is illustrated in Fig. 5. In this example it is evident that a single body part may carry significant information regarding the activity.

4 EXPERIMENTAL EVALUATION

4.1 Datasets

As we have already mentioned, our study has focused both on a) activities that resemble to "activities of daily living" (ADLs); and b) on "medical conditions." For training and experimental evaluation we relied on parts of two publicly available datasets. More specifically, for ADLs we have used the PKU-MMD dataset [16], while for medical conditions the NTU RGB+D [25] dataset. Both target to continuous multi-modal 3D HAR, providing RGB, depth, infrared and skeletal joint sequences for each activity. Activities have been captured using the Microsoft Kinect v2 sensor. From PKU-MMD we have selected 11 classes that are considered to be mostly related to ADLs: *eat meal snack*, *falling*, *handshaking*, *hugging other person*, *make a phone call answer phone*, *playing with phone tablet*, *reading*, *sitting down*, *standing up*, *typing on a keyboard* and *wearing a jacket*. Also, from NTU RGB+D we have selected the medical-condition-related category consisting of 12 classes, namely: *sneeze/cough*, *staggering*, *falling down*, *headache*, *chest pain*, *back pain*, *neck pain*, *nausea/vomiting*, *fan self*, *yawn*, *stretch oneself* and *blow nose*. The number of samples per class for both datasets is depicted in Table 1.

Table 1. Samples per class of the datasets used.

PKU-MMD			NTU-RGB+D		
class	**training**	**testing**	**class**	**training**	**testing**
eat meal/snack	381	42	*sneeze/cough*	865	83
falling	357	39	*staggering*	855	92
handshaking	189	21	*falling down*	856	90
hugging other person	189	21	*headache*	843	103
make a phone call/answer phone	308	34	*chest pain*	859	88
playing with phone/tablet	458	50	*back pain*	846	102
reading	387	42	*neck pain*	846	101
sitting down	496	55	*nausea/vomitting*	853	92
standing up	495	54	*fan self*	845	101
typing on a keyboard	387	42	*yawn*	860	97
wear jacket	411	45	*stretch oneself*	862	97
			blow nose	864	95
Total:	**4058**	**445**	**Total:**	**10254**	**1141**

4.2 Experimental Setup and Network Training

Experiments were performed on a personal workstation with an Intel™i7 5820K 12 core processor on 3.30 GHz and 16GB RAM, using NVIDIA™Geforce GTX 2060 GPU with 8 GB RAM and Ubuntu 18.04 (64 bit). The deep architecture has been implemented in Python, using Keras 2.2.4 [4] with the Tensorflow 1.12 [1] backend. All data pre-processing and processing steps have been implemented in Python 3.6 using NumPy, SciPy and OpenCV. For training, we used the ReLu activation function except from the last dense layer wherein softmax activation function was used. Moreover, we set the batch size to 32. We used the RMSprop optimizer, set the dropout to 0.5, set the learning rate to 0.001 and trained for 80 epochs. Moreover, since the duration of each activity was set to 150 frames, upon interpolation, the size of the input images was 25×149×3.

4.3 Results and Discussion

Firstly we evaluated the proposed methodology under the assumption of no occlusion. Experiments are depicted in Tables 2 and 3 for PKU-MMD and NTU-RGB+D, respectively. As it may be observed, mean F_1 score was 0.95 and 0.67, respectively. Then, we assessed the contribution of different body parts to the accuracy of classification. Intuitively one should expect 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.

In case of PKU-MMD, our experiments indicated that all parts were needed to maximize accuracy. When one leg is omitted, a small performance drop was observed; mean F_1 score was 0.91 and 0.90, upon the removal of left and right leg, respectively. Accordingly, when two legs are omitted, a further small performance drop was observed, compared to the previous case, i.e., leading to mean F_1 score equal to 0.75. However, as it was expected and has been experimentally verified, the omission of each arm led to a significant performance drop; mean F_1 score was equal to 0.70 and 0.61 for the left and right arm, respectively. Of course, upon removing both arms led to a mean F_1 score equal to 0.19 which is

insufficient for real-life applications. Finally, when one arm and one leg have been simultaneously removed, the mean F_1 scores were 0.66 and 0.58 for the left and the right side, respectively.

Upon careful observation of the confusion matrices depicted in Fig. 6, for each occlusion case we should notice the following, when comparing with the case where all joints had been used:

- *Left Arm*: upon observance of activity examples in the dataset, actors perform *handshaking* and *making a phone call* mainly using their left hand. When left arm is removed, a significant performance drop is observed mainly on those activities. Therefore, *handshaking* is misclassified as *hugging other person*, *sitting down* or *standing up*, while *making a phone call* is misclassified as *playing with phone/tablet*, *sitting down* or *standing up*. Moreover, the performance of *eat meal/snack* and *wear jacket* drop, although actors use both hands and misclassified in both cases as *hugging other person* or *reading*
- *Right Arm*: the loss of accuracy in the case that the right arm is missing from the skeleton is due to the fact that most of the actors performing actions in the dataset are right handed. In the case of class *playing with phone* which is misclassified as *making a phone call* or *sitting down*, the loss of accuracy is primarily caused by the similarity of the actions as well as the similarity in the posture of the rest of the skeleton structure during these actions. Moreover, *reading* and *typing on a keyboard* are both misclassified as *sitting down*, while the former is misclassified as *handshaking* and the latter as *falling*
- *Left & Right Arm*: the recognition of all activity classes is primarily based on the hand movements of the actors. When both arms are removed, the accuracy of all classes is significantly reduced as expected. Exceptions to the above statement are primarily the activities *sitting down*, *standing up* and secondarily *falling*, wherein the trained model predicts the activity correctly, regardless of arm movements
- *Left Leg*: removing the left leg of the actors does not significantly affect the action recognition process because most of the actions are based on hand-movement features. A small drop of performance is observed in *typing on keyboard*
- *Right Leg*: also in this case, removing the right leg of the actor does not significantly affect the action recognition, for the aforementioned reason. A small drop of performance is observed in *typing on keyboard* and *make a phone call/answer phone*
- *Left & Right Leg*: when both legs are removed from the skeleton, the network misclassifies *falling* as *sitting down* primarily because it detects the change of the actor's head level. Moreover, *typing on a keyboard* is misclassified as *playing with phone/tablet*. A smaller drop of performance is also observed in case of *hugging other person* which is misclassified as *wear jacket*
- *Left Arm & Left Leg*: the drop of performance compared to the case of left arm, is due to the fact that some actions that contain leg movements are also affected from the removal of both left limbs. Thus, *eat meal/snack* is misclassified as *hugging other person* or *reading*, *handshaking* as *hugging other person* or *sitting down*, *make a phone call/answer phone* as *playing with phone/tablet* or *standing up* and *wear jacket* as *reading* or *hugging other person*
- *Right Arm & Right Leg*: the drop of performance compared to the case of right arm, is due to the fact that some actions that contain leg movements are also affected from the removal of both right limbs. Thus, *eat meal/snack* is misclassified as *handshaking*, *falling* as *sitting down*, *hugging other person* as *handshaking*, *falling*, *handshaking* or *sitting down*, *make a phone call/answer phone* as *standing up*, *playing with phone/tablet* as *make a phone call/answer phone* and *typing on a keyboard* as *make a phone call/answer phone* and *standing up*.

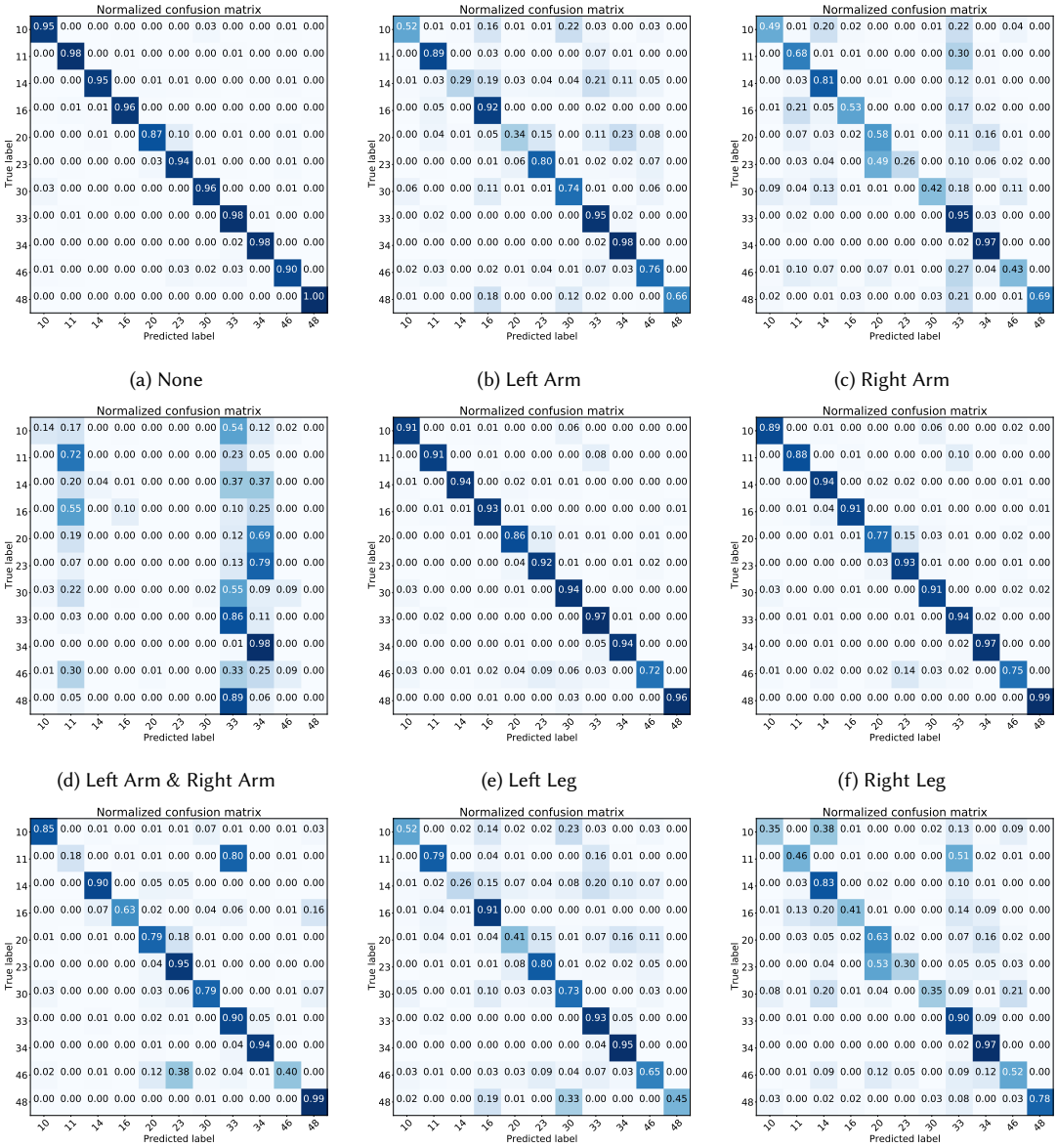
Table 2. Experimental results of the proposed approach for the 11 selected classes of PKU-MMD dataset. P, R, F_1 denote Precision, Recall, F_1 score, respectively. By "None" we denote the case wherein all body parts are included. LA, RA, LL, LR denote the occlusion of left arm, right arm, left leg, right leg, respectively. Classes are denoted as: 10:eat meal/snack, 11:falling, 14:handshaking, 16:hugging other person, 20:make a phone call/answer phone, 23:playing with phone/tablet, 30:reading, 33:sitting down, 34:standing up, 46:typing on a keyboard, 48:wear jacket.

	PKU-MMD																										
	None			**LA**			**RA**			**LA & RA**			**LL**			**RL**			**LL & RL**			**LA & LL**			**RA & RL**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
10	0.95	0.95	0.95	0.85	0.52	0.65	0.79	0.49	0.61	0.77	0.14	0.24	0.92	0.91	0.92	0.94	0.89	0.92	0.93	0.85	0.89	0.81	0.52	0.63	0.74	0.35	0.48
11	0.97	0.98	0.97	0.84	0.89	0.86	0.62	0.68	0.65	0.32	0.72	0.45	0.95	0.91	0.93	0.96	0.88	0.92	0.99	0.18	0.30	0.87	0.79	0.83	0.75	0.46	0.57
14	0.96	0.95	0.96	0.85	0.29	0.43	0.44	0.81	0.57	0.82	0.04	0.08	0.93	0.94	0.94	0.83	0.94	0.88	0.85	0.90	0.87	0.67	0.26	0.37	0.32	0.83	0.46
16	0.95	0.96	0.96	0.40	0.92	0.56	0.75	0.53	0.62	0.74	0.10	0.17	0.91	0.93	0.92	0.92	0.91	0.91	0.96	0.63	0.76	0.42	0.91	0.57	0.74	0.41	0.53
20	0.94	0.87	0.90	0.70	0.34	0.45	0.40	0.58	0.48	0.11	0.00	0.01	0.85	0.86	0.86	0.89	0.77	0.83	0.68	0.79	0.73	0.60	0.41	0.48	0.38	0.63	0.48
23	0.90	0.94	0.92	0.82	0.80	0.81	0.94	0.26	0.40	0.00	0.00	0.00	0.86	0.92	0.89	0.79	0.93	0.86	0.64	0.95	0.76	0.76	0.80	0.78	0.84	0.30	0.45
30	0.94	0.96	0.95	0.65	0.74	0.70	0.91	0.42	0.57	1.00	0.02	0.04	0.84	0.94	0.89	0.86	0.91	0.88	0.86	0.79	0.83	0.52	0.73	0.61	0.87	0.35	0.50
33	0.95	0.98	0.96	0.73	0.95	0.83	0.44	0.95	0.61	0.26	0.86	0.40	0.87	0.97	0.92	0.89	0.94	0.92	0.57	0.90	0.70	0.71	0.93	0.80	0.51	0.90	0.65
34	0.99	0.98	0.98	0.79	0.98	0.87	0.80	0.97	0.88	0.33	0.98	0.50	0.98	0.94	0.96	0.97	0.97	0.97	0.94	0.94	0.94	0.81	0.95	0.87	0.71	0.97	0.82
46	0.96	0.90	0.93	0.75	0.76	0.75	0.68	0.43	0.53	0.44	0.09	0.15	0.94	0.72	0.82	0.91	0.75	0.82	0.91	0.40	0.55	0.72	0.65	0.68	0.56	0.52	0.54
48	1.00	1.00	1.00	1.00	0.66	0.80	1.00	0.69	0.82	0.00	0.00	0.00	0.99	0.96	0.97	0.97	0.99	0.98	0.86	0.99	0.92	1.00	0.45	0.62	1.00	0.78	0.88
Mean	**0.96**	**0.95**	**0.95**	**0.76**	**0.71**	**0.70**	**0.71**	**0.62**	**0.61**	**0.44**	**0.27**	**0.19**	**0.91**	**0.91**	**0.91**	**0.90**	**0.90**	**0.90**	**0.84**	**0.76**	**0.75**	**0.72**	**0.67**	**0.66**	**0.67**	**0.59**	**0.58**

Moreover, in case of NTU-RGB+D, although in general drop of performance upon removal of body parts was similar to what has been observed in PKU-MMD, interestingly, upon removal of legs, mean F_1 score was increased to 0.76 and 0.75 for the cases of left and right leg, respectively and was almost equal to the full body case when both legs were removed, with a mean F_1 score equal to 0.66. Apart from that, upon removal of left and of right arm lead to mean F_1 scores equal to 0.45 and 0.59, respectively, while removal of both arms lead mean F_1 score equal 0.17, similar to the one of PKU-MMD in the same case. Finally, when one arm and one leg have been simultaneously removed, the mean F_1 scores were 0.45 and 0.54 for the left and the right side, respectively.

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:

- *Left Arm*: most of the classes under evaluation in NTU-RGB+D involve arm movements and are performed with the left arm and/or both arms. By removing the left arm, *sneeze/cough*, *headache*, *neck pain*, *fan self*, *yawn* and *blow nose* are misclassified as *chestpain* and *back pain*, while *stretch oneself* is misclassified as *yawn*
- *Right Arm*: more accurate than the previous case, since very few actions are heavily dependent on this arm of the actor so as to be distinguished from other classes. However, *stretch oneself* is misclassified mainly as *yawn*
- *Left & Right Arm*: By removing both arms, the model tries to classify the action based on the core body joints and misclassifies almost all actions as *chestpain*, thus performance is very low. However, *staggering* and *chestpain* do not exhibit a significant drop of performance
- *Left Leg*: minimal drop of accuracy is observed, since no actions are solely recognizable by left leg movement
- *Right Leg*: same case as left leg, i.e., minimal drop of accuracy is observed, since no actions are solely recognizable by right leg movement.
- *Left & Right Leg*: most of the classes are not affected by the removal of both legs because the hand movements are those that influence prediction. However, in case of *staggering* wherein both legs are equally important for the classification, the accuracy levels drop, as it is misclassified to the majority of other classes
- *Left Arm & Left Leg*: performance drop is much similar to the one of Left Arm. This leg does not have a significant influence in the classification process
- *Right Arm & Right Leg*: performance drop is much similar to the one of Right Arm. This leg does not have a significant influence in the classification process

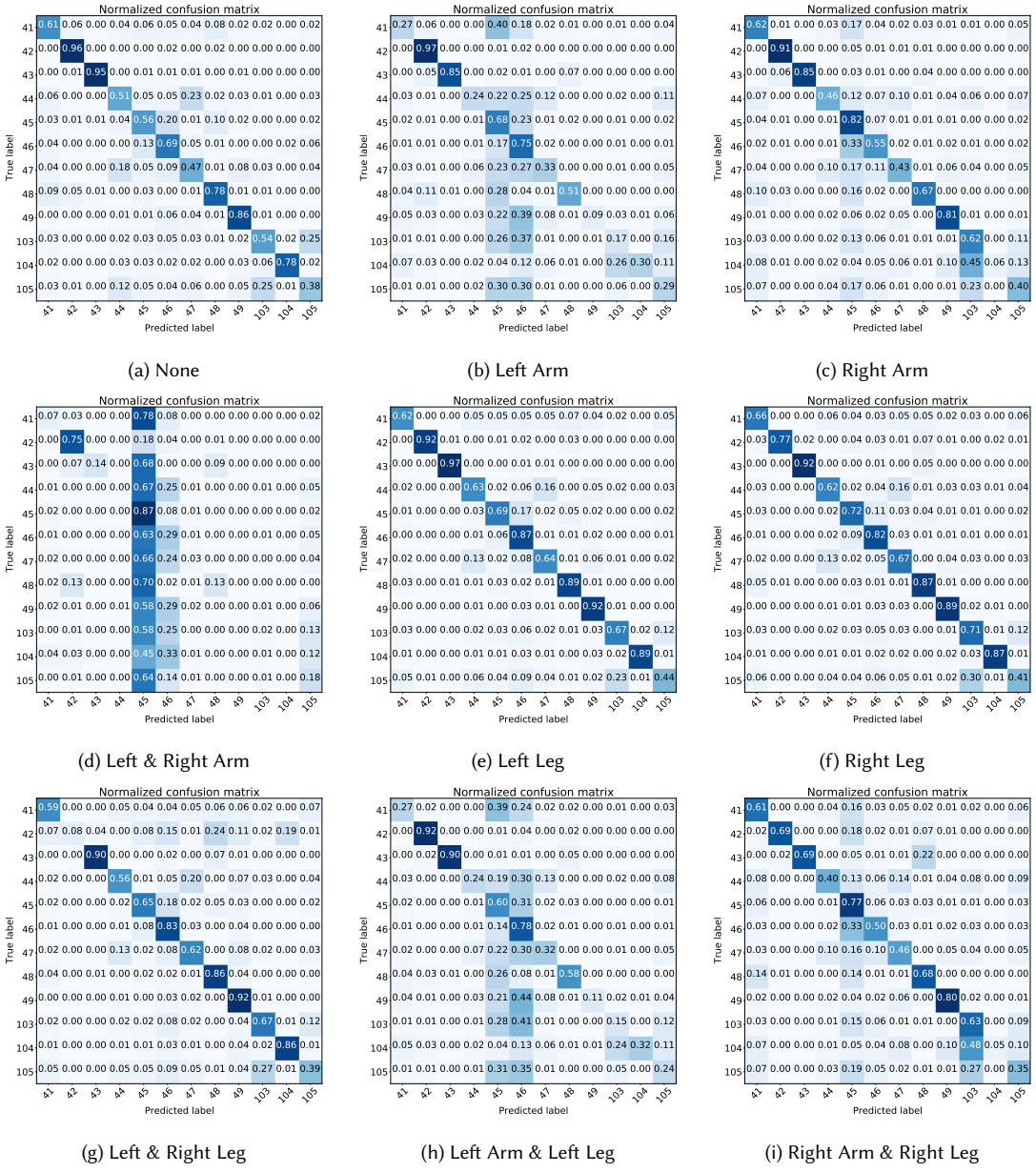


(a) None (b) Left Arm (c) Right Arm

(d) Left Arm & Right Arm (e) Left Leg (f) Right Leg

(g) Left Leg & Right Leg (h) Left Arm & Left Leg (i) Right Arm & Right Leg

Fig. 6. Normalized confusion matrices for recognition in the PKU-MMD data set (a) without removing any body part, (b)–(i) upon removing the body part(s) denoted in the caption of the corresponding subfigure. Classed depicted in matrices are: 10:eat meal/snack, 11:falling, 14:handshaking, 16:hugging other person, 20:make a phone call/answer phone, 23:playing with phone/tablet, 30:reading, 33:sitting down, 34:standing up, 46:typing on a keyboard, 48:wear jacket.



(a) None (b) Left Arm (c) Right Arm

(d) Left & Right Arm (e) Left Leg (f) Right Leg

(g) Left & Right Leg (h) Left Arm & Left Leg (i) Right Arm & Right Leg

Fig. 7. Normalized confusion matrices for recognition in the NTU-RGB+D data set (a) without removing any body part, (b)–(i) upon removing the body part(s) denoted in the caption of the corresponding subfigure. Classed depicted in matrices are: 41:sneeze/cough, 42:staggering, 43:falling down, 44:headache, 45:chest pain, 46:back pain, 47:neck pain, 48: nausea/vomiting, 49: fan self, 103:yawn, 104:stretch oneself, 105: blow nose.

Table 3. Experimental results of the proposed approach for the 12 selected classes of NTU RGB+D dataset. P, R, F_1 denote Precision, Recall, F_1 score, respectively. By "None" we denote the case wherein all body parts are included. LA, RA, LL, LR denote the occlusion of left arm, right arm, left leg, right leg, respectively. Classes are denoted as: 41:sneeze/cough, 42:staggering, 43:falling down, 44:headache, 45:chest pain, 46:back pain, 47:neck pain, 48: nausea/vomiting, 49: fan self, 103:yawn, 104:stretch oneself, 105: blow nose.

	NTU RGB+D																										
	None			**LA**			**RA**			**LA&RA**			**LL**			**RL**			**LL&RL**			**LA&LL**			**RA&RL**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
41	0.62	0.61	0.62	0.50	0.27	0.35	0.57	0.62	0.60	0.37	0.07	0.12	0.79	0.62	0.69	0.72	0.66	0.69	0.70	0.59	0.64	0.55	0.27	0.36	0.52	0.61	0.56
42	0.87	0.96	0.91	0.75	0.97	0.85	0.87	0.91	0.89	0.71	0.75	0.73	0.98	0.92	0.95	0.97	0.77	0.86	1.00	0.08	0.15	0.87	0.92	0.90	0.94	0.69	0.79
43	0.98	0.95	0.97	0.98	0.85	0.91	0.99	0.85	0.92	0.96	0.14	0.25	0.98	0.97	0.98	0.98	0.92	0.95	0.95	0.90	0.93	0.98	0.90	0.94	0.99	0.69	0.81
44	0.52	0.51	0.51	0.61	0.24	0.35	0.64	0.46	0.53	0.29	0.00	0.01	0.66	0.63	0.64	0.67	0.62	0.65	0.65	0.56	0.60	0.59	0.24	0.34	0.62	0.40	0.48
45	0.58	0.56	0.57	0.24	0.68	0.35	0.36	0.82	0.50	0.12	0.87	0.21	0.71	0.69	0.70	0.69	0.72	0.70	0.66	0.65	0.66	0.22	0.60	0.33	0.33	0.77	0.46
46	0.51	0.69	0.59	0.26	0.75	0.38	0.52	0.55	0.53	0.11	0.29	0.19	0.58	0.87	0.70	0.64	0.82	0.72	0.52	0.83	0.64	0.23	0.78	0.36	0.51	0.50	0.51
47	0.49	0.47	0.48	0.50	0.33	0.40	0.59	0.43	0.50	0.33	0.03	0.06	0.66	0.64	0.65	0.62	0.67	0.65	0.59	0.62	0.61	0.48	0.32	0.38	0.51	0.46	0.48
48	0.74	0.78	0.76	0.80	0.51	0.62	0.84	0.67	0.74	0.51	0.13	0.21	0.83	0.89	0.86	0.79	0.87	0.83	0.66	0.86	0.75	0.79	0.58	0.67	0.64	0.68	0.66
49	0.74	0.86	0.79	0.91	0.09	0.16	0.76	0.81	0.78	0.20	0.00	0.00	0.76	0.92	0.84	0.82	0.89	0.86	0.62	0.92	0.74	0.86	0.11	0.20	0.76	0.80	0.78
103	0.61	0.54	0.57	0.29	0.17	0.21	0.43	0.62	0.51	0.26	0.02	0.04	0.68	0.67	0.68	0.60	0.71	0.65	0.63	0.67	0.65	0.29	0.15	0.20	0.40	0.63	0.49
104	0.94	0.78	0.86	0.93	0.30	0.46	0.95	0.06	0.11	0.00	0.00	0.00	0.96	0.89	0.92	0.94	0.87	0.90	0.80	0.86	0.83	0.92	0.32	0.48	0.90	0.05	0.10
105	0.45	0.38	0.41	0.34	0.29	0.31	0.46	0.40	0.43	0.27	0.18	0.22	0.64	0.44	0.52	0.58	0.41	0.48	0.56	0.39	0.46	0.34	0.24	0.28	0.43	0.35	0.38
Mean	**0.67**	**0.67**	**0.67**	**0.59**	**0.45**	**0.45**	**0.67**	**0.60**	**0.59**	**0.34**	**0.21**	**0.17**	**0.77**	**0.76**	**0.76**	**0.75**	**0.74**	**0.75**	**0.70**	**0.66**	**0.64**	**0.59**	**0.45**	**0.45**	**0.63**	**0.55**	**0.54**

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented a study on the effect of occlusion in the context of a human activity recognition methodology. Our study focused on the recognition of activities of daily living and medical conditions and used two publicly available datasets. As baseline approach we used a convolutional neural network, whose input was a 2D representation of skeletal motion. Our goal was to assess how partial occlusion of the subject affected the accuracy of recognition and experimented with artificial occlusion of body parts, i.e., we removed the corresponding joints from the aforementioned representation from the entire activity. For activity recognition, we used the model that had been trained with the whole skeleton.

Our extensive experiments showed that the removal of arms had a significant effect on accuracy. This was not surprising, as most of the selected activities were expressed mainly by one or both arms' motion. Of course, in some cases several activities showed quite consistent performance, despite the removal of one or more body parts. However, in our opinion, as demonstrated, occlusion is one of the dominant problems in human activity recognition applications. Therefore, in the future we would like to incorporate occluded samples within the training process of our method.

6 ACKNOWLEDGMENTS

This project has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No 273 (Funding Decision:ΓΤΕΤ122785/I2/19-07-2018).

REFERENCES

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, ... & Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In 12th USENIX symposium on operating systems design and implementation (OSDI 16) (pp. 265-283).

[2] Angelini, F., Fu, Z., Long, Y., Shao, L., & Naqvi, S. M. (2019). 2d pose-based real-time human action recognition with occlusion-handling. IEEE Transactions on Multimedia, 22(6), 1433-1446.

[3] Berretti, S., Daoudi, M., Turaga, P., & Basu, A. (2018). Representation, analysis, and recognition of 3D humans: A survey. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(1s), 16.

[4] Chollet, F. (2015) *Keras*, https://github.com/fchollet/keras.

[5] Debes, C., Merentitis, A., Sukhanov, S., Niessen, M., Frangiadakis, N., & Bauer, A. (2016). Monitoring activities of daily living in smart homes: Understanding human behavior. IEEE Signal Processing Magazine, 33(2), 81-94.

[6] Du, Y., Fu, Y., & Wang, L. (2015, November). Skeleton based action recognition with convolutional neural network. In 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR) (pp. 579-583). IEEE.

[7] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE.

[8] Gu, R., Wang, G., & Hwang, J. N. (2020). Exploring Severe Occlusion: Multi-Person 3D Pose Estimation with Gated Convolution. arXiv preprint arXiv:2011.00184.

[9] Hou, Y., Li, Z., Wang, P., & Li, W. (2018). Skeleton optical spectra-based action recognition using convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology, 28(3), 807-811.

[10] Iosifidis, A., Tefas, A., & Pitas, I. (2012). Multi-view human action recognition under occlusion based on fuzzy distances and neural networks. In 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO) (pp. 1129-1133). IEEE.

[11] Ke, Q., An, S., Bennamoun, M., Sohel, F., & Boussaid, F. (2017). Skeletonnet: Mining deep part features for 3-d action recognition. IEEE signal processing letters, 24(6), 731-735.

[12] Keogh, A., Dorn, J. F., Walsh, L., Calvo, F., & Caulfield, B. (2020). Comparing the Usability and Acceptability of Wearable Sensors Among Older Irish Adults in a Real-World Context: Observational Study. JMIR mHealth and uHealth, 8(4), e15704.

[13] Kinect SDK. Published online at Microsoft.com. Retrieved from https://developer.microsoft.com/en-us/windows/kinect/, 06/03/2021.

[14] Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: self-maintaining and instrumental activities of daily living. The gerontologist, 9(3 Part 1), 179-186.

[15] Liu, T., Sun, J. J., Zhao, L., Zhao, J., Yuan, L., Wang, Y., ... & Adam, H. (2020). View-Invariant, Occlusion-Robust Probabilistic Embedding for Human Pose. arXiv preprint arXiv:2010.13321.

[16] Liu, C., Hu, Y., Li, Y., Song, S., & Liu, J. (2017). PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. arXiv preprint arXiv:1703.07475.

[17] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[18] Li, C., Hou, Y., Wang, P., & Li, W. (2017). Joint distance maps based action recognition with convolutional neural networks. IEEE Signal Processing Letters, 24(5), 624-628.

[19] Liu, M., Liu, H., & Chen, C. (2017). Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition, 68, 346-362.

[20] Long-Term Care in America: Expectations and Preferences for Care and Caregiving. Published online at longtermcarepoll.org. Retrieved from https://www.longtermcarepoll.org/long-term-care-in-america-expectations-and-preferences-for-care-and-caregiving/, 06/03/2021.

[21] Majumder, S., Mondal, T., & Deen, M. J. (2017). Wearable sensors for remote health monitoring. Sensors, 17(1), 130.

[22] Papadakis, A., Mathe, E., Vernikos, I., Maniatis, A., Spyrou, E.,& Mylonas, P. (2019). Recognizing human actions using 3d skeletal information and CNNs. In Int'l Conf. on Engineering Applications of Neural Networks. Springer, Cham.

[23] Rashidi, P., & Mihailidis, A. (2012). A survey on ambient-assisted living tools for older adults. IEEE journal of biomedical and health informatics, 17(3), 579-590.

[24] Roser, M., Ortiz-Ospina, E., & Ritchie H. (2013). Life Expectancy. Published online at OurWorldInData.org. Retrieved from: https://ourworldindata.org/life-expectancy, 06/03/2021.

[25] Shahroudy, A., Liu, J., Ng, T. T., & Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1010-1019).

[26] United Nations. (2020) World Population Ageing 2020 Highlights.

[27] Vernikos, I., Mathe, E., Papadakis, A., Spyrou, E., & Mylonas, P. (2019). An image representation of skeletal data for action recognition using convolutional neural networks. In Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (pp. 325-326).

[28] Wang, P., Li, W., Ogunbona, P., Wan, J., & Escalera, S. (2018). RGB-D-based human motion recognition with deep learning: A survey. Computer Vision and Image Understanding, 171, 118-139.

[29] Wang, P., Li, W., Li, C., & Hou, Y. (2018). Action recognition based on joint trajectory maps with convolutional neural networks. Knowledge-Based Systems, 158, 43-53.